VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV TELEKOMUNIKACÍ
DEPARTMENT OF TELECOMMUNICATIONS

ANALÝZA HOSTOVÁNÍ WEBOVÝCH SERVERŮ
ANALYSIS OF WEB SERVER HOSTING

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. Petr Ilgner**
AUTHOR

VEDOUCÍ PRÁCE — **doc. Ing. Dan Komosný, Ph.D.**
SUPERVISOR

BRNO 2017

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

magisterský navazující studijní obor **Telekomunikační a informační technika**
Ústav telekomunikací

Student: Bc. Petr Ilgner ***ID:*** 154739
Ročník: 2 ***Akademický rok:*** 2016/17

NÁZEV TÉMATU:

Analýza hostování webových serverů

POKYNY PRO VYPRACOVÁNÍ:

Naprogramujte aplikaci pro detekci hostování webových stránek na pronajatém serveru (web hosting) u poskytovatelů v ČR. Vstupem aplikace budou zvolené webové stránky (doménová jména) registrované v ČR. Výstupem bude odhad, kolik webových stránek je umístěno na pronajatých serverech. Odhad proveďte pomocí analýzy IP adres a doménových jmen. Aplikaci sestavte v programovacím jazyce Python.

DOPORUČENÁ LITERATURA:

[1] PUŽMANOVÁ, R. TCP/IP v kostce. 2. vyd. Kopp, 2009. 620 s. ISBN: 978-80-7232-388-3.

[2] PILGRIM, M. Ponořme se do Python(u) 3. CZ.NIC, 2010. 435 s. ISBN: 978-80-904248-2-1.

Termín zadání: 1.2.2017 ***Termín odevzdání:*** 24.5.2017

Vedoucí práce: doc. Ing. Dan Komosný, Ph.D.
Konzultant:

doc. Ing. Jiří Mišurec, CSc.
předseda oborové rady

ABSTRAKT

Tato diplomová práce se zabývá problematikou hostování webových stránek v České republice. Cílem práce je navrhnout postup a na jeho základě realizovat aplikaci pro automatizované určení, zda je konkrétní web provozován na serveru třetí strany nebo na vlastním serveru organizace či jednotlivce.

Práce je rozdělena na tři hlavní části, kde první, teoretická část pojednává o principech, protokolech a službách, které jsou potřebné pro zajištění dostupnosti webových stránek s přihlédnutím na specifika prostředí českého internetu. V další, praktické části, je na základě získaných teoretických znalostí navrženo několik přístupů k analýze získaných dat. Na jejich základě je navržen a formou detekční aplikace implementován algoritmus, jehož cílem je určit, zda jsou předložené vstupní webové stránky umístěné na sdíleném serveru a pro každou z nich shromáždit informace o jejím hostování. V poslední části práce je aplikace spuštěna na vytvořené databázi webových stránek dle kategorií subjektů provozující tyto weby. V této, analytické části, jsou výstupní získané informace o hostování těchto webů zpracovány a prezentovány spolu s uvedením příslušných souvislostí týkajících se hostování webů v České Republice.

KLÍČOVÁ SLOVA

protokol, web, hosting, webhosting, doména, serverhosting, virtualní server, data mining, analýza

ABSTRACT

Master thesis deals with the problematics of webpages hosting in the Czech Republic. The goal of the thesis is to design a procedure and to implement an application for automated determination of whether a particular website is operated on a third party server or an own server of the organization or the individual.

The thesis is divided into two main parts. The first part of the thesis lays a theoretical base concerning principals, protocols and services which are necessary for providing availability of the webpages with special attention to the specifics of the Czech internet environment. In the practical part there are several approaches to the analysis of the obtained data proposed. On that basis there is designed and in the form of the detection application implemented an algorithm which goal is to determine whether the submitted webpages are placed on a shared server and to collect the information about their hosting. In the last part of the thesis the application runs based on the created database of webservers organized by the categories of the webserver operators. In this analytic part the obtained output information about these webservers hosting are processed and presented together with the relevant context concerning the hosting of these webpages in the Czech Republic.

KEYWORDS

protocol, web, hosting, webhosting, domain, serverhosting, virtual server, data mining, analysis

ILGNER, Petr *Analýza hostování webových serverů*: diplomová práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, Rok. 89 s. Vedoucí práce byl doc. Ing. Dan Komosný, Ph.D.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma „Analýza hostování webových serverů“ jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autora

PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu diplomové práce, doc. Ing. Danovi Komosnému, Ph.D., za jeho příkladné odborné vedení, poskytování cenných konzultací a podnětné návrhy k práci.

Poděkování patří také Bohu, rodině a přátelům, které mi dal a kteří mě trpělivě podporovali při zpracování této práce.

Brno ....

podpis autora

SIX research centre
sensor, information and communication systems

Faculty of Electrical Engineering
and Communication

Brno University of Technology
Technicka 12, CZ-61600 Brno, Czechia

http://www.six.feec.vutbr.cz

Výzkum popsaný v této diplomové práci byl realizovaný v laboratořích podpořených projektem Centrum senzorických, informačních a komunikačních systémů (SIX); registrační číslo CZ.1.05/2.1.00/03.0072, operačního programu Výzkum a vývoj pro inovace.

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

EVROPSKÁ UNIE
EVROPSKÝ FOND PRO REGIONÁLNÍ ROZVOJ
INVESTICE DO VAŠÍ BUDOUCNOSTI

2007–13
OP Výzkum a vývoj pro inovace

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

SEZNAM VÝPISŮ

ÚVOD

Počátky sítě Internet, na které jsou provozovány webové servery sahají do sedmdesátých let dvacátého století [7]. Ze sítě původně zamýšlené pro vojenské a výzkumné použití se stala síť, která je součástí našich životů. Internet se stal pro mnohé primárním zdrojem pro získávání informací, univerzálním komunikačním prostředkem, i zábavní platformou. Ze sítě, jejíž uživatelé byli nadšení experti pečující o její rozvoj, se Internet stal sítí, kterou používají obyčejní lidé bez hlubších technických znalostí a povědomí o složitosti a komplexitě této sítě.

S tím, jak se síť Internet z výzkumné a později univerzitní sítě otevřel dalším komerčním subjektům a v důsledku toho i privátní sféře, s měnící se podobou a schopností koncových zařízení, se mění i dominantní služby, které jsou na této síti používané. První použitou službou, na které byla také demonstrována funkčnost sítě, se stala elektronická pošta. [1] Ačkoliv je e-mail stále stěžejní a velmi důležitou komunikační službou, pro mnoho uživatelů se stala prakticky synonymem internetu nikoliv poštovní, ale webová služba.

Web byl k dispozici od doby, kdy se připojení k internetu stalo dostupné i mimo univerzity a právě na této službě, webu, je zřejmé, jak se její stránky proměnili. S rostoucími schopnostmi počítačů a dalších zařízení, které ji využívají, přibyl nejen datově náročný multimediální obsah, ale změnila se také nenávratně filozofie samotných webů. Webové stránky v dnešní době nejsou pouhými neinteraktivními dokumenty svázanými prostými hypertextovými odkazy, ale mnoho služeb je interaktivních a lze sledovat, jak se web stále častěji stává vstupní branou k dalším službám Internetu.

Uživatelé chtějí čerpat z výhod, které z připojení k celosvětové síti plynou, bez toho, aby museli mít zevrubné technické znalosti. Jen těžko by si například elektronická pošta získala tolik uživatelů a tak univerzální použití, pokud by bylo nutné pro každého uživatele vlastnit speciální server, ten následně bez výpadků provozovat, nakonfigurovat na něm síťový operační systém, v něm poštovního agenta, a průběžně zajišťovat požadovanou úroveň bezpečnosti. Až díky komerčním poskytovatelům, kteří za úplatu nebo zdarma poskytují na své infrastruktuře prostor pro e-mailovou schránku nebo vlastní webovou prezentaci, se tyto služby staly široké veřejnosti dostupné.

Tato práce se soustředí na problematiku umístění webových zdrojů. Cílem práce je navrhnout postup a na jeho základě realizovat aplikaci, která bude automatizovaně rozhodovat, zda je daná webová prezentace provozována na serveru třetí strany nebo na vlastním serveru organizace či jednotlivce.

V první kapitole práce je poskytnut úvod do problematiky hostování webových serverů, jsou vysvětleny základní principy a služby, se kterými je dále v rámci práce

nakládáno. Zvláštní pozornost je pak kladena konkrétní situaci v českém internetovém prostředí a organizacím participujícím na provozu internetových domén a webhostingových služeb. V následující kapitole je představena navržená metodika detekce hostování webových stránek, jsou popsány čtyři dílčí detekční metody a algoritmus pro stanovení celkového výsledku na základě vyhodnocení dílčích metod. Následující kapitola popisuje realizovanou aplikace umožňující automatizovaný sběr dat a následné vyhodnocení těchto údajů pro vstupní seznam webů podle navržených detekční metod. Poslední kapitola je tvořena analýzou dat, které autor získal použitím realizovaných nástrojů pro různé kategorie institucí provozující webové stránky, a která si klade za cíl na tomto vzorku analyzovat specifika hostování webů v České republice.

Realizovaný nástroj lze při dostatečné a aktuální množině vstupních webů použít jako nástroj pro hodnocení poskytovatelů při výběru webhostingu. Získaná data z aplikace dokáží poskytnout informace o množství webů hostovaných na jednotlivých serverech, jejich umístění a podpoře různých technologií.

Protože problematika hostování webových stránek je velmi rozsáhlá a mnoho pojmů není zcela jednoznačných, v úvodu práce jsou definované některé dále používané pojmy. U pojmů uvedených v tabulce č. 1 je v práci implicitně uvažován tabulkou popsaný význam daných pojmů, pokud není řečeno jinak.

Tab. 1: Definice používaných pojmů

Pojem	**Význam pojmu**
Webový server	Soubor hardwarových a softwarových prostředků určených pro doručení požadovaných hypertextových dokumentů. Blíže specifikován v kapitole 1.5.
Webová stránka	Konkrétní hypertextový dokument umístěný na webovém serveru.
Webové stránky	Soubor hypertextových dokumentů, běžně označované jako „web“, přístupné typicky pod určitým doménovým jménem. Pojem je blíže popsán v kapitole 1.1.
Hostování webových stránek	Umístění webových stránek na webovém serveru.
Hostované webové stránky	Webové stránky, které nejsou umístěny na vyhrazeném serveru pro tyto stránky a jejím příslušejícím webům. Tyto sdílené webové servery jsou provozovány a spravovány jiným subjektem než konkrétní na nich umístěné weby. Příklad některých webů a rozhodnutí, zda jsou hostované, je uveden v tabulce č. 2.
Webhoster	Poskytovatel hostingových služeb webových stránek. Více v kapitole 1.8.

Tab. 2: Vzorové scénáře hostování webových stránek

Web	Vlastník domény	Webový server	Vlastník serveru	Hostovaná
www.vutbr.cz	VUT v Brně	piranha.ro.vutbr.cz	VUT v Brně	NE
eprihlaska.vutbr.cz	VUT v Brně	piranha.ro.vutbr.cz	VUT v Brně	NE
www.ilgner.cz	Petr Ilgner	80.79.28.101	Petr Ilgner	NE
www.albert.cz	AHOLD CR, a.s.	178.238.35.100	AHOLD CR, a.s.	NE
www.zelenakocka.cz	ViKa servis s.r.o.	wl26-f185.wedos.net	WEDOS Internet, a.s.	ANO
www.petrsrna.cz	Petr Srna	wl26-f185.wedos.net	WEDOS Internet, a.s.	ANO

1 PROBLEMATIKA HOSTOVÁNÍ WEBOVÝCH STRÁNEK

1.1 Webové stránky

Službou World Wide Web (WWW) rozumíme systém na sobě nezávislých serverů, které poskytují uživatelům webové stránky (hypertextové dokumenty) na základě požadavků zaslaných z jejich webových prohlížečů.

Tyto hypertextové dokumenty jsou mezi sebou propojeny odkazy, které mohou vést na jiné hypertextové dokumenty, ale i jiné služby Internetu. Uživatel pak celý systém vnímá interaktivně, tedy může kliknutím na určité prvky (textové odkazy, obrázky nebo další navigační prvky) webové stránky přejít na jinou podstránku nebo jiný web [7].

Tyto webové stránky, objekty WWW systému, jsou soustředěny ve webovém místě na webovém serveru (webovém místě), často nazývaném pouze jako „web“. Webové, ale i další zdroje, jsou identifikovány pomocí identifikátoru URL.

Protokolem pro vyžádání konkrétního webového zdroje a jeho odeslání do prohlížeče je Hypertext Transfer Protocol (HTTP) a jeho nadstavba Hypertext Transfer Protocol Secure (HTTPS), která umožňuje kryptografické zabezpečení spojení mezi klientem a serverem i ověření identity protistrany.

1.2 Identifikátor URL a identifikace umístění

Uniform Resource Locator (URL) je řetězcem, který přesně identifikuje umístění zdrojů informací na webu. Častokrát je nazýván pojmem „webová adresa“. URL je specifickým typem Uniform Resource Identifier (URI) Jeho struktura je přesně definovaná [2]. Syntaxe každého URL koreluje se obecnou syntaxí URI, jejíž obecná forma je:

`scheme:[//[user:password@]host[:port]][/]path[?query][#fragment]`

Skládá se z:

- **schématu**, které specifikuje použitý protokol. Každé schéma by mělo být registrováno organizací Internet Assigned Numbers Authority (IANA) a dále je v URL následováno dvojtečkou. Pro webové zdroje se setkáváme se schématy `http:` a `https:`. Některá schémata dále vyžadují uvedení dvou lomítek `//`.
- **autoritativní části**, která volitelně specifikuje uživatelské jméno a heslo nutné pro přístup ke zdroji, adresy serveru (hosta), která je tvořena doménovým jménem případně IP adresou, přičemž Internet Protocol version 4 (IPv4) adresa musí být uvedena ve tvaru decimálních čísel oddělených tečkou a Internet

Protocol version 6 (IPv6) adresa musí být uzavřena v hranatých závorkách. Volitelně může následovat dvojtečka a číslo portu. Pokud není číslo portu uvedeno, uvažuje se obvyklý port pro dané schéma, v případě protokolu HTTP je to port 80, pro protokol zabezpečený protokol HTTPS je to 443.

- **cesty**, která obsahuje řetězec obvykle organizovaný v hierarchické podobě a jeho části jsou odděleny lomítkem („/“). Často tato cesta odpovídá relativní cestě systému souborů webového serveru vztaženou k kořenovému adresáři, není to ale pravidlem a záleží na implementaci. Z pohledu přístupnosti je vhodné, aby měla cesta člověkem dobře čitelný a pochopitelný tvar.
- volitelného **dotazu** oddělený od cesty otazníkem („?“) sestávajícího se z řetězce určujících specifických dotazů na daný zdroj. Na webu se nejčastěji používá forma atribut–hodnota oddělená oddělovacím znakem „&“.
  Například. `atribut1=hodnota1&atribut2=hodnota2`.
- volitelného **doplňujícího fragmentu** oddělený od předešlé části znakem „#“. Ten poskytuje odkaz na příslušnou sekci v dokumentu. Pokud je zdrojem Hypertext Markup Language (HTML) dokument, přejde na element, který odpovídá identifikátoru ve fragmentu.

Kromě výše popsaného celého, absolutního, způsobu adresování lze odkazy uvádět také **relativně**, celá adresa se pak odvodí od toho, kde se nachází zdroj na který odkazujeme. Jazyk HTML dovoluje pomocí tagu `base` změnit adresu, která se bude na stránce považovat jako referenční pro relativní adresy.
Příklad použití: `<base href="http://www.feec.vutbr.cz">`.

Je definováno také protokolově relativní URL, (označováno PRL nebo PRURL), která nemá specifikován cílový protokol [15]. Pro `//vutbr.cz` se použije protokol zdroje, ze kterého se na příslušnou stránku odkazujeme (typicky tedy HTTP nebo HTTPS).

1.3 Protokol HTTP

Protokol transferu hypertextových informací HTTP je aplikační protokol pro distribuované, spolupracující informační systémy, které používají hypermédia. Jde o obecný, bezstavový protokol, který může být použit pro přístup ke zdrojům různého charakteru a přenos informací [5].

Protokol HTTP je založen na zprávách typu požadavek a odpověď, pracuje v režimu klient-server. Spolu s elektronickou poštou je nejvíce používaným aplikačním protokolem na internetu [13].

Nejpoužívanější verze HTTP/1.1 byla definována v RFC 2616 [5]. V roce 2014 byl tento dokument nahrazen více RFC 7230-7237 [6] samostatně definující syntaxi

zpráv, jejich sémantiku a obsah, tvary požadavků kešování a autentizaci v rámci protokolu HTTP/1.1.

Podstatnou inovací protokolu HTTP/1.1 je zasílání jména hosta (domény nebo IP adresy) v HTTP požadavcích. Díky tomu je možné, aby jeden webový server (míněno softwarová instance) obsluhovala více webových míst.

Pomocí WWW klienta uživatel zašle WWW serveru požadavek v textové formě (plain-text), která obsahuje identifikaci požadovaného zdroje a informace o schopnostech prohlížeče. Server poté odpovídá datovou větou, která se skládá z hlavičky, ve které popisuje výsledek dotazu, identifikuje své softwarové vybavení, sděluje webovému prohlížeči důležité metainformace o způsobu nakládání s odpovědí, použitém kódování, datu a času vyřízení dotazu. Za touto hlavičkou v případě úspěšného požadavku následují data samotného požadovaného dokumentu. Webový server se může rozhodnout, jakou verzi dokumentu odešle na základě datových informací, které klient uvádí v požadavku [15].

Protokol HTTP je svým charakterem bezstavový, konkrétní požadavek není svázán s konkrétní odpovědí. Tento protokol tedy neumí uchovávat stav a kontext komunikace a nelze určit, zda dva požadavky od jednoho zdroje spolu souvisí. Vzhledem k tomu, že web často dovoluje uživatelům určitou interaktivitu, kdy navrácené odpovědi mohou být dynamicky generovány na základě vstupních argumentů, ale i uživatelském kontextu, byl HTTP protokol rozšířen o technologie sessions a cookies, které umožňují konkrétnímu webovému serveru uchovat určité informace o stavu spojení na počítači uživatele.

Příklad HTTP požadavku a příslušné odpovědi je zachycen na výpisu 1.1 a 1.2.

1.4 Protokol HTTPS

Protokol HTTPS využívá protokol HTTP spolu s protokoly Secure Sockets Layer (SSL) a Transport Layer Security (TLS). Zajišťuje zabezpečenou komunikace webového prohlížeče s webovým klientem, kdy je pomocí asymetrické kryptografie ověřena identita serveru, případně i klienta [7].

Pro tyto kryptografické protokoly je důležitá infrastruktura veřejných klíčů Public Key Infrastructure (PKI), která předpokládá důvěru v zaslaný certifikát, který bývá podepsán jednou z certifikačních autorit, její certifikáty jsou uloženy v úložišti důvěryhodných certifikátů operačního systému nebo prohlížeče.

V roce 2014 vznikla nekomerční certifikační autorita Let's Encrypt, která zavedla automatizovaný proces získávání a obnovování certifikátů pro hostované webové stránky. Certifikační autorita splnila podmínky většiny programů pro uznání důvěryhodnosti certifikačních autorit operačních systémů a také její certifikáty byly

podepsány certifikačními autoritami, které jsou obecně považovány za důvěryhodné. Díky bezplatnému a jednoduchému procesu vydávání certifikátů se stala autorita značně oblíbená a její vznik rozšířil globální využití protokolu HTTPS [9].

Výpis 1.1: Příklad HTTP požadavku.

```
Host: www.feec.vutbr.cz
User-Agent: Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:50.0) Gecko
   /20100101 Firefox/50.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q
   =0.8
Accept-Language: en-US,en;q=0.5
Accept-Encoding: gzip, deflate
Referer: https://www.google.cz/
Cookie: PHPSESSID=vhcphapsr5d75am985r3hvu9v5;
Connection: keep-alive
Upgrade-Insecure-Requests: 1
Cache-Control: max-age=0
```

Výpis 1.2: Příklad hlavičky HTTP odpovědi.

```
Cache-Control: no-store, no-cache, must-revalidate, post-check=0,
   pre-check=0
Connection: Keep-Alive
Content-Language: en
Content-Location: home.php.en
Content-Type: text/html; charset=ISO-8859-2
Date: Thu, 24 Nov 2016 12:48:45 GMT
Expires: Thu, 19 Nov 1981 08:52:00 GMT
Keep-Alive: timeout=5, max=98
Pragma: no-cache
Server: Apache/1.3.42 Ben-SSL/1.59 (Unix) PHP/5.6.27
TCN: choice
Transfer-Encoding: chunked
Vary: negotiate
X-Powered-By: PHP/5.6.27
```

1.5 Webový server

Jako webový server (WWW nebo HTTP server) je v této práci uvažován soubor hardwarových a softwarových prostředků potřebných k vyřizování HTTP požadavků od klientů. Obvykle jde o odeslání cíle, který je specifikován URL v konkrétním HTTP požadavku.

Z hlediska softwarových prostředků se jedná o počítačový program, který v souladu s protokolem HTTP provádí tyto činnosti. Tento program je provozován pod operačním systémem spuštěným ať již virtualizovaně nebo přímo na hardwarovém serveru.

Nejpoužívanějšími webovými servery jsou v současné době Apache, nginx a Internet Information Server (IIS) [13].

Z hlediska hardwarových prostředků se v drtivé většině jedná o dedikované servery, které jsou uzpůsobené k nepřetržitému provozování softwarového webového serveru, případně dalšího softwarového vybavení pro jedno nebo více webových míst (prezentací).

Aby mohl být zaslán klientský požadavek na konkrétní webový server v síti Internet, je nutné, aby byly servery v celosvětové síti Internet adresovatelné. Každému serveru, který má být dostupný v internetové síti, musí být přidělena internetová adresa, tzv. IP adresa. Jedná se o číselný identifikátor, který jednoznačně identifikuje síťová rozhraní v počítačových sítích využívající internetového protokolu IP. Nejrozšířenější je využití protokolu IP verze 4 (IPv4), který používá 32bitové adresy. Tyto adresy jsou typicky zapsány dekadicky po jednotlivých oktetech.

IP adresy jsou děleny na **veřejné** a **privátní**. Veškeré tyto adresy jsou přidělovány žadatelům organizací IANA. Některé bloky IP adres byly ovšem vyhrazeny pro privátní použití, tyto adresy jsou nazývány jako privátní nebo neveřejné.

Organizace IANA pro přidělování a registraci IP adres využívá regionální registry Resource Internet registry (RIR). Každý tento RIR přiděluje IP adresy pro svou konkrétní oblast. Adresy protokolu IPv4 jsou přidělovány jednotlivým RIR po velkých blocích, obvykle po 2^{24} adresách.

Pro oblast Evropy, Blízkého východu a centrální Asie je registrátorem organizace Réseaux IP Européens Network Coordination Centre (RIPE NNC) [8].

1.6 Systém DNS

Každý uzel podporující TCP/IP má svou IP adresu, kterou je v síti identifikován. Vzhledem k tomu, že se tento údaj špatně pamatuje, je praktičtější a pro uživatele příjemnější používat jména než číselné adresy. Systém DNS zajišťuje překlad doménových jmen (které označují i webové místo) na konkrétní platnou IPv4 adresu o délce 32 bitů nebo IPv6 adresu o délce 128 bitů.

Z pohledu problematiky webových serverů, musí mít každý webový server, který chce komunikovat, přidělenou vlastní IP adresu. Aby byla stránka dobře zapamatovatelná a také přenositelná mezi různými webovými servery, zavádí se používání

doménových jmen pro webové servery. Vzhledem k tomu, že veškerou webovou komunikaci vždy iniciuje uživatel, není nutné, aby měl také klient přiděleno unikátní doménové jméno.

1.6.1 Doménové jméno

Proto, aby nebyli uživatelé nuceni pamatovat si IP adresy serverů, byl realizován systém Domain Name System (DNS), který pomocí hierarchie DNS serverů umožňuje provádět vzájemné překlady doménových jmen a IP adres uzlů sítě. Díky postupnému rozšiřování k dalším účelům dnes slouží coby distribuovaná databáze síťových informací.

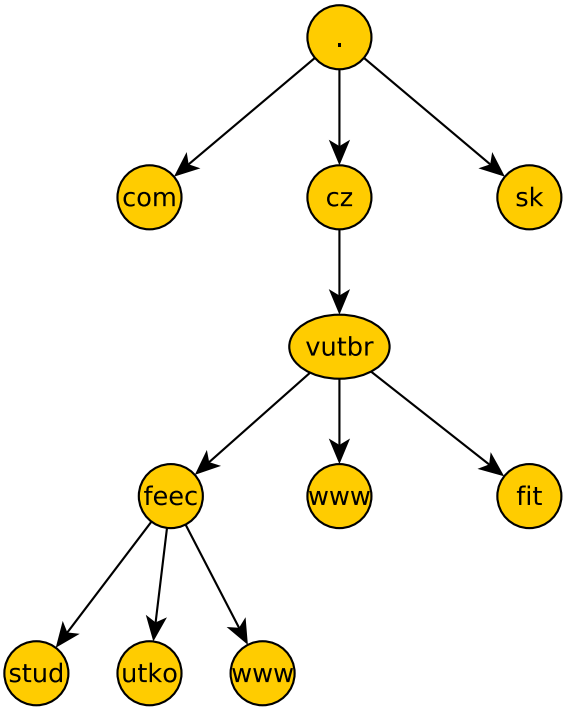
Celé doménové jméno se skládá z několika částí, které jsou odděleny tečkami a celé toto jméno končí tečkou. Část nejvíce vpravo je doménou nejvyšší úrovně – Top Level Domain (TLD), směrem doleva se nachází domény vyššího řádu, přičemž závěrečná tečka se běžně vynechává.

Příkladem doménového jména je `fest.stud.feec.vutbr.cz`, který označuje server „fest“ umístěný ve studentské doméně Fakulty Elektrotechniky a komunikačních technologií Vysokého učeního technického v Brně, umístěného v české národní TLD.

1.6.2 Hierarchie domén

Vzhledem k velikosti databáze doménových jmen není možné a žádoucí, aby byl celý systém spravován centralizovaně. DNS je založen na decentralizovaném systému jmenných domén, které mají stromovou reprezentaci. Strom může mít až 128 úrovní, značené od nuly (tj. kořen stromu) až k listům stromu, které mají maximální úroveň 127. Každý uzel stromu má svoji značku, která je tvořena řetězcem o maximální délce 63 znaků. Kořen má nulovou značku – jde tedy o prázdný řetězec.

Jako Fully Qualified Domain Name (FQDN) je označováno kompletní jméno uzlu, včetně tečky pro nulovou značku kořene, které se běžně nepoužívá.



Obr. 1.1: Příklad hierarchické organizace systému DNS.

Příklad hierarchického stromu jmen, kde v listu grafu najdeme mj. doménové jméno `stud.feec.vutbr.cz.` je zobrazen na obrázku č. 1.6.2.

Vrchol stromu tvoří „kořenová doména“ (doména nulté úrovně), další úroveň (doména první úrovně) (tedy TLD) tvoří domény nejvyšší úrovně, pod nimi jsou domény druhé, třetí a další úrovně.

Domény vrcholové (TLD) úrovně (například cz, com, edu) se dělí na:

- generické domény - třípísmenné a
- národní domény - dvoupísmenné a to podle mezinárodních kódů států definovaných normou ISO 3166, které označují příslušnost k dané zemi. Správcem v dané zemi je pověřený registr.

1.6.3 Národní doména .cz

Před rozdělením Československa v roce 1993 byla přidělena doména .CS, v roce 1994 byla československá národní doména první doménou, která byla zrušena [7]. Ve stejném roce byla přidělena pro Českou republiku doména .CZ. Správcem národní domény České republiky je sdružení CZ.NIC.

CZ.NIC je zájmové sdružení právnických osob a provozuje centrální registr, tedy databázi o Doménových jménech, jejich držitelích a dalších osobách vedená sdružením CZ.NIC. Centrální registr je zdrojem pro delegaci Doménových jmen do zóny CZ vedené primárním jmenným serverem [14].

Tato organizace ale nenabízí registraci domén konkrétním koncovým osobám. Domény je možné registrovat prostřednictvím akreditovaných registrátorů. Tyto komerční subjekty nabízejí služby registrace a vedení domén.

Pravidla registrace doménových vydané sdružením CZ.NIC specifikují požadavky na tvorbu doménových jmen:

- může obsahovat 1-63 znaků,
- může obsahovat znaky a-z a číslice 0-9,
- nesmí začínat a končit znakem pomlčky („-“),
- nesmí obsahovat dvě pomlčky za sebou [14].

Majitelem domény se může stát jakákoliv fyzická a právnická osoba. Doménové jméno je možné registrovat na jeden rok i více let, nejvýše však na 10 let.

Registr domén v zóně .cz provozuje CZ.NIC na adrese `https://www.nic.cz/whois/`, případně prostřednictvím protokolu WHOIS na `whois.nic.cz`.

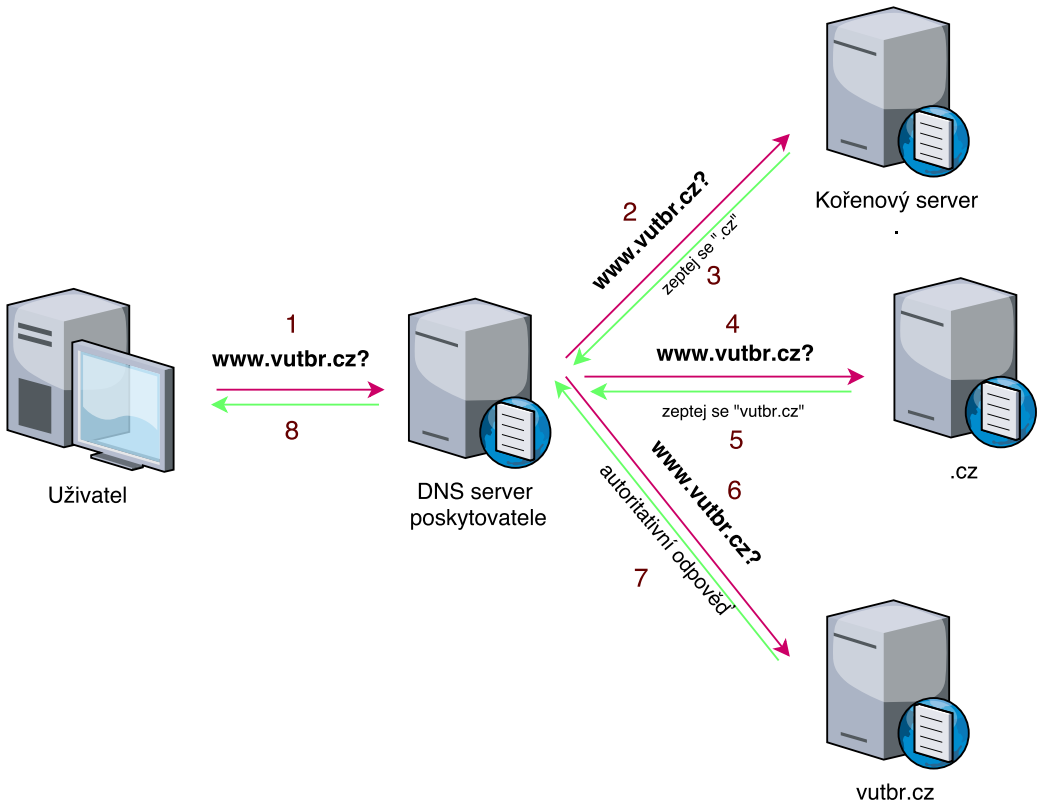
1.6.4 Autoritativní rekurzivní DNS servery

V DNS systému rozlišujeme autoritativní a rekurzivní DNS servery. Při posílání dotazu si navíc může klient zvolit, zda požaduje provést tzv. rekurzivní dotaz. V tom případě, nezná-li dotazovaný DNS server sám odpověď (tedy není autoritativní pro dotazovanou doménu), spustí algoritmus pro vyhledávání odpovědi, který začíná se u kořenových DNS serverů, na základě jejich odpovědí se dotazuje konkrétnější DNS serverů. Rekurzivní servery jsou kešovací, po získání odpovědi na dotaz si ji uchovávají do lokální keše a v případě opakovaného požadavku po dobu platnosti Time To Live (TTL) odpoví uloženou informací z této keše.

Autoritativní DNS servery publikují tu část DNS stromu, která byla jmennému serveru přidělena. Prakticky řečeno, autoritativním servery jsou například kořenové servery domény nejvyšší úrovně nebo DNS servery konkrétní domény 2. úrovně. Rekurzivními servery pak jsou DNS servery, které poskytují Internet Service Provider (ISP) svým zákazníkům a DNS klienti uživatelů internetu se na ně dotazují.

V případě, že budeme po svém rekurzivním DNS serveru požadovat překlad domény `www.vutbr.cz`, nejprve musí rekurzivní DNS server zjistit, jaké mají DNS servery CZ.NIC. Tento dotaz nejprve provede u jednoho ze třinácti jmenných serverů [15]. Teprve poté co zjistí jejich adresy, pošle na primární DNS server této zóny DNS dotaz, kdy zjistí adresy jmenných serverů Vysokého učeního technického v Brně. Teprve poté konečně pošle další dotaz, kterým zjistí požadovaný dotaz.

Princip rekurzivního dotazu ilustruje obrázek 1.6.4.

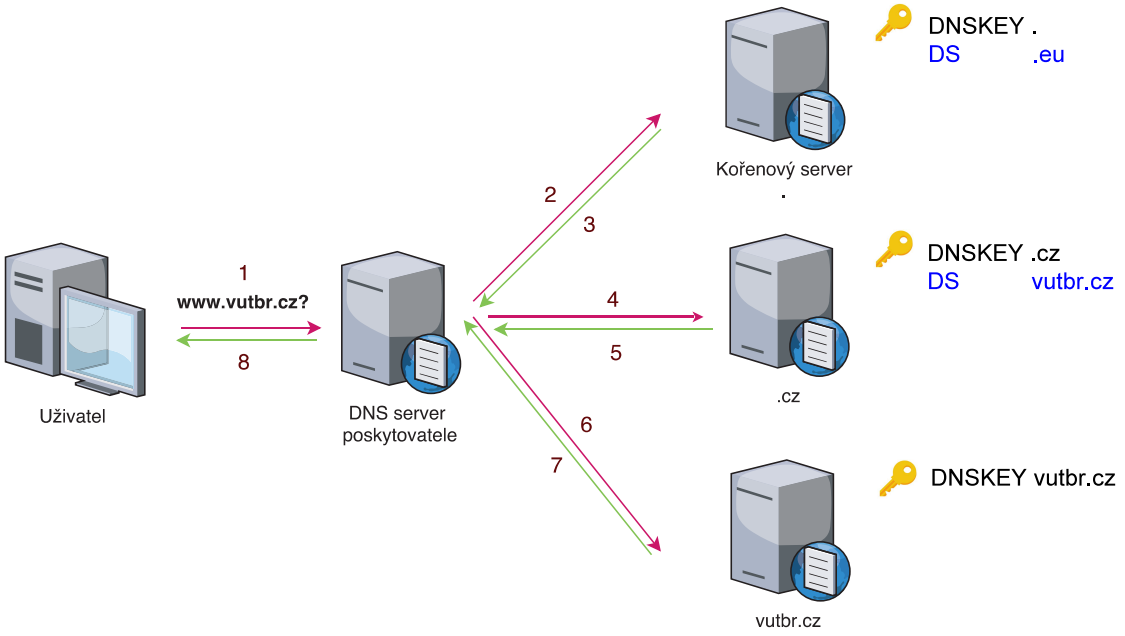


Obr. 1.2: Princip rekurzivních dotazů v DNS systému.

1.6.5 Systém DNSSEC

Systém DNSSEC zajišťuje rozšíření systému DNSSEC o prvky asymetrické kryptografie, tak aby bylo zajištěno ověření autenticity zjištěných údajů ze systému DNS tak, aby nebylo možné v rámci nešifrovaného protokolu DNS uživateli podvrhnout neplatnou odpověď DNS serveru a tím uživatele nasměrovat na jiný server, než se kterým zamýšlel komunikovat.

Držitel domény vygeneruje dvojici soukromého a veřejného klíče. Svým soukromým klíčem podepíše údaje, které do systému DNS vkládá. Pomocí veřejného klíče je pak možné zpětně ověřit pravost tohoto podpisu. Aby byl tento veřejný klíč dostupný všem, je publikován u nadřazeného DNS serveru, tento nadřazený server má svůj veřejný klíč publikovaný u svého nadřazeného serveru a dále [10]. Tím je vytvořen tzv. řetěz důvěry, který zajišťuje důvěryhodnost údajů v systému DNS. Příklad vytvoření takového řetězu pro doménu vutbr.cz je zobrazen na obrázku 1.6.5., ve kterém jsou ikonou klíče vyznačeny soukromé klíče.



Obr. 1.3: Řetěz důvěry v systému DNSSEC.

V případě, že není DNSSEC implementován na cílové doméně nebo DNS resolveru, který užívá uživatel, nejsou odpovědi z DNS serveru důvěryhodné. Zajištění důvěryhodnosti je řešeno také protokolem HTTPS, protože certifikáty jsou vystaveny pro konkrétní doménové jméno a certifikační autorita je povinna ověřit, zda je žadatel oprávněn vystavit certifikát pro danou doménu.

Technologie DNSSEC byla v národní .cz doméně spuštěna v roce 2008 [11].

1.7 Virtuální hosting

Virtuální hosting je metoda pro hostování více webových míst (doménových jmen) na jednom serveru. Tento jeden server sdílí prostředky jako jsou procesorové cykly, operační paměť bez toho, aby bylo nutné mít vyhrazeny zvláštní prostředky pro každou webovou stránku.

Toho se široce využívá pro sdílený webový hosting, který umožňuje provozovat vlastní webovou prezentaci za nižší cenu, než jaké jsou náklady při hostování každých webových stránek na vlastním serveru.

Virtuální hosting založený na doménových jménech

Díky rozšíření hlavičky HTTP protokolu o pole „Host“ je možné rozlišit, ke které webové stránce se požadavek vztahuje. Tedy na jedné adrese a jednom portu je možné provozovat více webových míst.

Příkladně pro požadavky směřující na doménu `www.feec.vutbr.cz` bude webový server považovat za kořenovou složku `/var/www/feec` a pro doménu `www.fit.vutbr.cz` bude uvažována složka `/var/www/fit`. Nejznámější implementací je modul apache-vhost webového serveru Apache [12].

Příklad konfigurace virtuálního hostingu na serveru Apache:

Výpis 1.3: Příklad konfigurace virtuálního hostingu webového serveru založeném na doménových jménech

```
Listen 80
<VirtualHost *:80>
DocumentRoot "/www/example1"
ServerName www.example.com
</VirtualHost>

<VirtualHost *:80>
DocumentRoot "/www/example2"
ServerName www.example.org
</VirtualHost>
```

U protokolu HTTPS bylo zavedeno rozšíření SNI (Server Name Indication), které ovšem nepodporují některé starší verze operačních systémů a prohlížečů.

Virtuální hosting založený na IP adresách

V tomto případě je systémem DNS každé webové místo směrováno na zvláštní IP adresu. Sdílený webový server má přiděleny na svém síťovém rozhraní všechny tyto IP adresy.

Webový server (software) nemusí podporovat žádné rozšíření pro virtuální hosting, postačí spustit více jeho instancích naslouchající na více IP adresách. Výhodou tohoto řešení je určitá vyšší nezávislost, nevýhodou je naopak nutnost více IP adres, což se vzhledem k vyčerpání IPv4 adres, vyplatí spíše pro větší projekty. U protokolu HTTPS je funkčnost zajištěna i na starších operačních systémech.

Příklad konfigurace virtuálního hostingu založeného na IP adresách na webovém serveru Apache je zobrazen ve výpisu 1.4.

Výpis 1.4: Příklad konfigurace virtuálního hostingu webového serveru Apache založené na IP adresách

```
Listen 80

<VirtualHost 172.20.30.40>
DocumentRoot "/www/example1"
ServerName www.example.com
</VirtualHost>

<VirtualHost 172.20.30.50>
DocumentRoot "/www/example2"
ServerName www.example.org
</VirtualHost>
```

Virtuální hosting založený na portových číslech

Využívá se toho, že kromě běžného čísla portu služby HTTP 80, resp. portu 443 pro protokol HTTPS, je možné webový server provozovat na jiných portech. Podobně jako v předchozím případě, běží více instancí webového serveru, pro každý port může běžet dokonce zcela jiný software. Nevýhodou je ale nutnost uvádět číslo portu u webové adresy, navíc z principu není možné zajistit navigaci z různých doménových na různé porty. Nestandardní porty jsou také mnohdy blokovány bezpečnostními politikami firewallů. Z těchto důvodů je tento způsob používán spíše pro vývoj nebo aplikace provozované v intranetovém prostředí.

1.8 Webhostingová služba

Proto, aby byla vytvořená webová stránka přístupná ostatním uživatelům odkudkoliv a kdykoliv, je nutné ji vystavit na webovém serveru. Jak již bylo naznačeno, aby nebylo nutné kvůli této jediné stránce pořídit nebo pronajmout specializovaný server, připojit jej k Internetu a spravovat, je vhodné pro menší až střední webové projekty využít služeb společností, které zajistí vystavení webové stránky na serverech připojených k Internetu vysokorychlostním připojením. Stránky pak budou umístěny na společném serveru spolu s dalšími weby. Taková služba se nazývá webhostingovou službou (nebo jen webhostingem) a její poskytovatel je nazýván webhosterem.

Výhodou je, že veškeré náklady na zakoupení odpovídajícího hardwaru, jeho provozování, energetické náklady, konektivitu serverů zajišťuje poskytovatel webhostingových služeb. V jeho režii je také následná údržba hardwaru, instalace a konfigurace softwarového vybavení, zajištění vyhovující bezpečnosti a dostupnosti.

Jako nevýhodu lze považovat určitou omezenou možnost konfigurace webového serveru, skriptovacího vybavení. Na serveru jsou aplikovány limity využívání zdrojů tak, aby se co nejvíce eliminovalo vzájemné ovlivnění provozovaných aplikací. Je nutné vzít v potaz, že veškerá data, která jsou umístěna na serveru webhostera jím mohou být získána a případně zneužita.

1.8.1 Rozdělení webhostingů

Klasický webhosting je nejrozšířenější formou hostingu webových stránek na internetu [17]. Na českém trhu je řada důvěryhodných společností, které nabízí kvalitní webhostingové služby s měsíčním poplatkem v řádu desítek až stovek korun. Součástí webhostingu většinou bývá kromě hostování webových stránek také mailhosting a vedení DNS záznamů. Při výběru konkrétního webhostingového programu je možné se rozhodovat podle různých parametrů, snadno srovnatelné jsou například:

- **Podporované technologie** – v případě, že webová prezentace není čistě statická, požadujeme podporu některé skriptovací technologie (nejčastěji PHP, ASP, Python, Ruby, Java).
- **Databáze** – pro rychlejší a sofistikovanější práci s daty jsou často využívány relační systémy řízení báze dat. Poskytovatel může nabízet jednu nebo více databázi pro využití z webové aplikace.
- **Diskový prostor** – určuje jak velký prostor nabízí zákazníkovi pro uložení svých dat (tj. webové prezentace a její zdroje a další soubory webové aplikace). Pokud je poskytnut prostor pro webové stránky a e-maily, je často počítán mimo tento limit.

- **Poštovní služby** – k většině hostingů je nabízena také možnost využívání jedné nebo více e-mailových schránek.
- **Maximální měsíční přenos dat** – někteří webhosteři jasně definují maximální množství přenesených dat k internetovým uživatelům. Pokud není takový limit stanoven, lze očekávat, že v případech nadměrného využívání přikročí k určitému omezení i „neomezené“ hostingy. V podmínkách je toto často označováno jako Fair Usage Policy (FUP).
- **Zálohování** – některé webhostingy nabízí v ceně služby automatizované zálohování webového obsahu včetně databází a e-mailových schránek. Někdy bývá účtován zvláštní poplatek za obnovení dat administrátory.
- **Protokolem pro správu dat** - většina hostingů nabízí pro nahrávání dat webové prezentace protokol File Transfer Protocol (FTP), resp. File Transfer Protocol Secure (FTPS). Někteří nabízí přístup ke konzoli operačního systému protokolem Secure Shell (SSH) a pro přenos dat protokol Secure Copy (SCP). Lze se setkat také s podporou různých verzovacích systémů (např. GIT nebo SVN).

Kromě webového serveru je samozřejmě nutné mít registrovanou doménu, na které bude webová stránka provozována. Někteří poskytovatelé nabízí možnost spuštění služby na doméně třetí úrovně v rámci jejich domény. Drtivá většina poskytovatelů nabízí při objednávce webhostingových služeb také registraci domény.

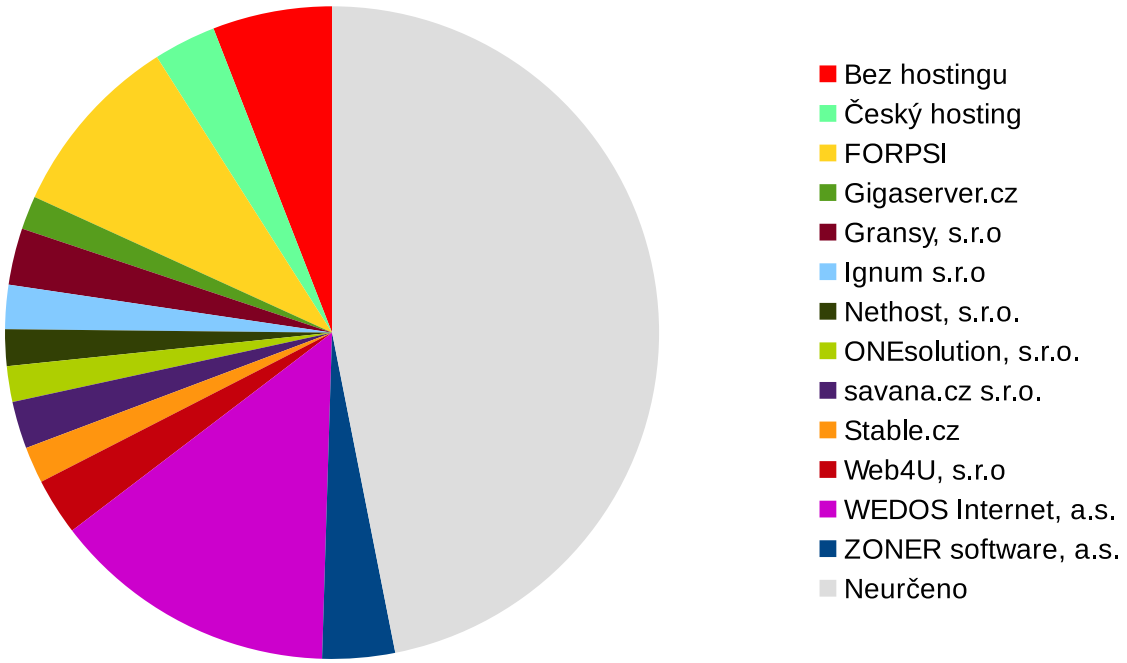
Zvláštním druhem webhostingů jsou služby, které nenabízí nahrávání dat žádným protokolem, ale poskytují pouze redakční systém, kterým je možné jednoduchým způsobem, nejčastěji prostřednictvím editoru typu What you see is what you get (WYSIWYG). Při jeho použití nemusí osoba spravující web umět používat značkovací jazyk, webový obsah je tvořen interaktivně v tomto redakčním systému. Takové redakční systémy je možné samozřejmě nahrát i na konvenční hostingy. Toto zjednodušené řešení ale upřednostní uživatelé, kteří nepožadují individuální konfiguraci provozovaného softwaru.

1.8.2 Webhosting v České republice

České národní domény .cz je možné samozřejmě hostovat také v zahraničí. Výhodou umístění obsahu na servery umístěné v České republice je rychlejší odezva načítání webových stránek pro zdejší uživatele, než v případě, kdy by byl web umístěn například na jiném kontinentě. Zahraniční konektivita mnoha sítí je také často méně dimenzovaná než připojení do národních sítí.

Správce národní domény, organizace CZ.NIC, zveřejňuje každý měsíc statistiku největších poskytovatelů webhostingu podle počtu hostovaných domén.

Tato data jsou získávána ze seznamu webových domén v .cz zóně, které má



Obr. 1.4: Statistika počtu domén hostovaných u jednotlivých společností v říjnu 2016 podle sdružení CZ.NIC. Zpracováno dle [17].

národní registrátor k dispozici. Příslušnost IP adresy, která je vedena v DNS záznamech ke konkrétnímu webhosterovi je primárně určena podle WHOIS záznamů v registru RIPE NCC a pro rozhodnutí se používá A záznam www nebo případně prázdný. (www.domena.cz nebo domena.cz).

Jako „bez hostingu“ jsou označeny weby, které nemají žádnou webovou stránku. Mohou to být nefunkční domény případně domény na kterých není provozován web. Jako „neurčeno“ jsou označeny ty domény, ke kterým se podle popsané metodiky organizace CZ.NIC nepodařilo přiřadit konkrétního webhostera nebo jsou provozovány na vyhrazených serverech.

1.9 Vyhrazené servery

V případě, že má webová stránka nebo organizace vlastní webový server, hovoříme o vyhrazeném serveru. Toto řešení je určeno zájemcům, kteří z například z výše uvedených nevýhod webhostingové služby nemohou pro umístění svých webových stránek použít sdílené servery.

Toto řešení je určené spíše odborníkům a podle míry, jaká část technického zajištění je outsourcována specializovanou firmou, rozlišujeme varianty vyhrazených serverů.

1.9.1 Provoz ve vlastním prostředí

Provozovatel webových stránek provozuje webový server ve své vlastní infrastruktuře, například v serverově organizace nebo univerzity. Vlastní tedy veškerou firemní infrastrukturu, veškerý hardware a zajišťuje kompletní provoz. K této variantě se nejčastěji rozhodují střední až velké organizace a klade největší požadavky na technické znalosti.

1.9.2 Serverhousing

V tomto případě provozovatel webové stránky vlastní fyzický server, který umístí do specializované serverovny poskytovatele, který za úplatu zajišťuje konektivitu pro server, napájení serveru a přístup k technickému vybavení zákazníka dle sjednaných podmínek. V tomto případě zákazník vlastní svůj server, veškerou okolní infrastrukturu ale provozuje a také vlastní poskytovatel služby. K serveru má zákazník plný vzdálený i fyzický přístup.

1.9.3 Dedikovaný server

V případě dedikovaného serveru poskytovatel zákazníkovi pronajímá fyzický server, který je určen pouze jemu. Zákazník tedy může na serveru provozovat cokoliv a celý výkon má k dispozici pouze pro sebe. Poskytovatel tedy vlastní veškerý hardware a je zodpovědný za jeho funkčnost, softwarové vybavení je ovšem plně v režii zákazníka. K serveru má plný vzdálený (administrátorský) přístup.

1.9.4 Spravovaný dedikovaný server

Na rozdíl od dedikovaného serveru poskytovatel zákazníkovi kromě pronájmu zařízení poskytuje správu serveru dle sjednaných podmínek. Tedy zodpovídá jak za funkčnost hardwaru, tak za funkčnost softwaru. Tento typ poskytované služby je v praxi často označován jako managed server.

1.9.5 Virtuální server

Virtuální privátní server (VPS) kombinuje výhody některých těchto přístupů a spolu výrazně nižší cenou může být vhodným řešením pro mnoho náročnějších webových serverů a projektů. Zákazník obdrží od poskytovatele „svůj“ server, na kterém může provozovat libovolný operační server a programové vybavení a tento server je logicky nezávislý. VPS ale sdílí fyzické prostředky serveru s dalšími zákazníky. To znamená, že na jednom fyzickém serveru jsou provozovány další virtuální privátní servery. Na základě sjednaných podmínek mohou být některé zdroje (například procesorová jádra, operační paměť) zákaznickému VPS smluvně přidělena. Podobně jako u webhostingu zákazník nevlastní nic, přesto je flexibilita řešení podobná plnohodnotnému vyhrazenému serveru a provozovatel VPS musí disponovat stejnou technickou znalostí jako u dedikovaného serveru. VPS má přidělenou vlastní veřejnou internetovou IP adresu.

2 NAVRŽENÁ METODIKA DETEKCE

Pro určení, zda je webová stránka umístěna na sdíleném serveru poskytovatele wehostingových služeb, je možné použít informace o webových serverech, které byly popsány v předchozí kapitole zabývající se problematikou hostování webových stránek.

Pro provádění rozhodování je nutné o každé doméně, která je přítomna ve vstupním souboru, který má být výslednou aplikací zpracován, získat co největší množství informací, které se dají pro automatizovanou detekci využít.

Neexistuje ovšem jednoznačný způsob, jak na základě těchto informací vždy zcela správně rozhodnou, zda je stránka hostovaná. Lze ovšem definovat několik přístupů, jaká data z sesbírané množiny dat budou užity k konečnému rozhodnutí a jak bude s těmito daty nakládáno. V rámci práce bylo navrženo několik detekčních metod:

1. Ověření shodnosti reverzních záznamů domény
2. Porovnání s databází známých webhosterů
3. Shodnost uvedeného držitele domény a sítě
4. Porovnání kontaktní adresy správce sítě

Tyto metody jsou dále popsány v následujících podkapitolách. V závěru kapitoly je navržen postup, kterým lze výsledky metod sumarizovat a určit tak celkový výsledek detekce.

2.1 Metoda 1 - Ověření shodnosti reverzních záznamů domény

U serverů, které jsou ve vlastnictví a správě určité organizace je očekávané, že správce prostřednictvím DNS záznamů nastaví PTR záznamy tak, aby IP adresa byla přeložitelná na doménové jméno serveru. Vytvoření PTR záznamů je nezbytné zejména pro odesílání e-mailů, protože velká část nastavení poštovních agentů požaduje, aby byla zdrojová IP adresa přijatých dopisů přeložitelná na doménové jméno a toto doménové jméno ukazovalo na danou IP adresu.

V případě, že název zkoumané domény odpovídá nalezenému PTR záznamu, lze konstatovat, že daný web není hostován na sdíleném serveru. Tato metoda nevrátí relevantní informace v případě, že není reverzní záznam korektně nastaven nebo daná organizace používá PTR záznamy, které nastavil poskytovatel připojení pro své jednotlivé IP adresy.

Pro ruční zjištění výsledků lze použít např. linuxový program `dig`. Pro přeložení adresy webu na IP adresu je možné využít sekvenci příkazu ve výpisu 2.1. Rozhodující data pro vyhodnocení metody (značeny červeně) pro webové stránky Mendelovy

univerzity v Brně uvádí tabulka 2.1.

Výpis 2.1: Zjištění výsledku reverzního DNS dotazu.

```
$ dig www.mendelu.cz AAA +short
valar.mendelu.cz.
195.178.72.2
$ dig +short -x 195.178.72.2
valar.mendelu.cz.
```

Tab. 2.1: Příklad vyhodnocení Metody č. 1.

Adresa webu	www.mendelu.cz
IP adresa serveru	195.178.72.2
Doménové jméno pro IP adresu	valar.mendelu.cz

2.2 Metoda 2 - Porovnání s databází známých webhosterů

V případě, že doménové jméno DNS záznamu ukazující na IP adresu zjištěného serveru je nalezeno v databázi známých webhostingů, považuje se takový web za hostovaný. Aplikace uchovává seznam těchto webhostingů v tabulce `webhoster` relační databáze.

Tato metoda nemusí zahrnout všechny poskytovatele hostingových služeb, navíc v případě, kdy je testována webová stránka samotného webhostera, bude metoda vyhodnocovat stránku jako hostovanou, což nemusí být pravda. Neplatné výsledky bude metoda vykazovat také v případě, že bude web hostován na dedikovaném serveru u poskytovatele, který poskytuje také webhostingové služby.

Pro ruční ověření výsledku této metody lze opět využít nástroj `dig` a dle sekvence příkazů uvedené ve výpisu 2.1 přeložit zjištěnou adresu webového serveru na doménové jméno a to následně porovnat s databází sdílených hostingů. Rozhodující data pro vyhodnocení metody (značeny červeně) pro webové stránky Ústavu zdravotnických informací a statistiky České republiky uvádí tabulka 2.2.

Tab. 2.2: Příklad vyhodnocení Metody č. 2.

Adresa webu	www.uzis.cz
IP adresa serveru	178.238.37.157
Doménové jméno pro IP adresu	yivo.onebit.cz
Nalezený webhoster	onebit.cz

2.3 Metoda 3 - Shodnost uvedeného držitele domény a sítě

V případě, že název organizace, která je držitelem doménového jména dle registru sdružení CZ.NIC, je podobný s názvem organizace, které je přidělen adresní prostor sítě, ve které se nachází IP adresa webserveru, považuje se web za nehostovaný.

Příklad vyhodnocení této metody pro stránky Českého vysokého učení technického v Praze je zobrazen v tabulce 2.3 (rozhodující údaje jsou vyznačeny červeně). Metoda zkoumá podobnost obou zjištěných řetězců, před porovnáním jsou řetězce upraveny, například je odstraněna přípona právní formy organizace, která se vyskytuje v registru domén, ale v registru přidělených sítí není uvedena.

Tab. 2.3: Příklad vyhodnocení Metody č. 3.

Adresa webu	www.cvut.cz
IP adresa serveru	147.32.3.202
Doménové jméno pro IP adresu	drupal7-prod.is.cvut.cz
Držitel domény	České vysoké učení technické v Praze
IP přidělena organizaci	Ceske vysoke uceni technicke v Praze

Ruční ověření výsledků této metody je možné porovnáním údajů v obou registrech, například programem `whois`. Příklad sekvence příkazů pro toto ověření je zobrazen ve výpisu 2.2.

Výpis 2.2: Porovnání informací z databáze WHOIS

```
contact:        SB:R1S-CES-8079-FA
org:            České vysoké učení technické v Praze
name:           České vysoké učení technické v Praze
address:        Zikova 4
address:        Praha 6
address:        16636
address:        CZ
e-mail:         neuman@vc.cvut.cz
(výpis zkrácen)
$ dig www.cvut.cz AAA +short
cvut.cz.
147.32.3.202
$ whois 147.32.3.202
organisation:   ORG-CVUT1-RIPE
org-name:       Ceske vysoke uceni technicke v Praze
address:        Ceske vysoke uceni technicke v Praze
address:        Zikova 1903/4
address:        Praha 6
address:        166 36
address:        The Czech Republic
abuse-mailbox:  abuse@cvut.cz
(výpis zkrácen)
```

Tato metoda může selhat, v případě, že držitel IP adresy není koncová organizace, ale poskytovatel připojení. Přímo přidělené IPv4 adresy mají zpravidla jen větší organizace. U této metody se provádějí dotazy pouze IPv4 adresy.

2.4 Metoda 4 - Porovnání kontaktní adresy správce sítě

Velké organizace mají často přidělený svůj adresní prostor v jejich registru přidělených sítí je možné najít e-mailový kontakt na držitele dané domény. Tento kontakt je často nazýván jako tzv. abuse kontakt, který má být využit pro nahlášení závadného provozu pocházejícího z daných sítí.

V případě, že byl nalezen tento kontakt pro danou IP adresu a doména této e-mailové adresy je shodná s doménou daného webu, lze tvrdit, že daný web je hostován přímo danou organizací.

Příklad údajů nutných k vyhodnocení metody pro webové stránky Českého statistického úřadu je v tabulce 2.4. Rozhodující údaje jsou zde vyznačeny červenou barvou.

Tab. 2.4: Příklad vyhodnocení Metody č. 4.

Adresa webu	czso.cz
IP adresa serveru	194.48.241.132
Přidělený adresní rozsah	194.48.241.0 - 194.48.241.255
Kontaktní e-mail sítě	jiri.lejnar@czso.cz

Ruční verifikaci lze provést podobně jako u předchozí metody, například pomocí programu `whois`, pomocí příkazů uvedených ve výpisu 2.2.

2.5 Určení celkového výsledku

Vzhledem k tomu, že ke každé metodě existují případy, kdy její rozhodnutí nebude správné, je nutné k určení, zda je stránka hostována na sdíleném serveru, použít výsledek všech metod. Zejména empirickým pozorováním byl zvolen aditivní přístup, kdy výsledek každé metody, která vrací odpověď, že je stránka hostována, zvyšuje skóre, které znamená, že stránka je hotován na webovém serveru, metody, které prokazují opak naopak toto skóre snižují. Princip určení celkového výsledku je zobrazen na obrázku 2.1.

Výsledné skóre je tedy určeno podle rovnice 2.5.

$$Skóre = Výsledek_1 + Výsledek_2 + Výsledek_3 + Výsledek_4 \tag{2.1}$$

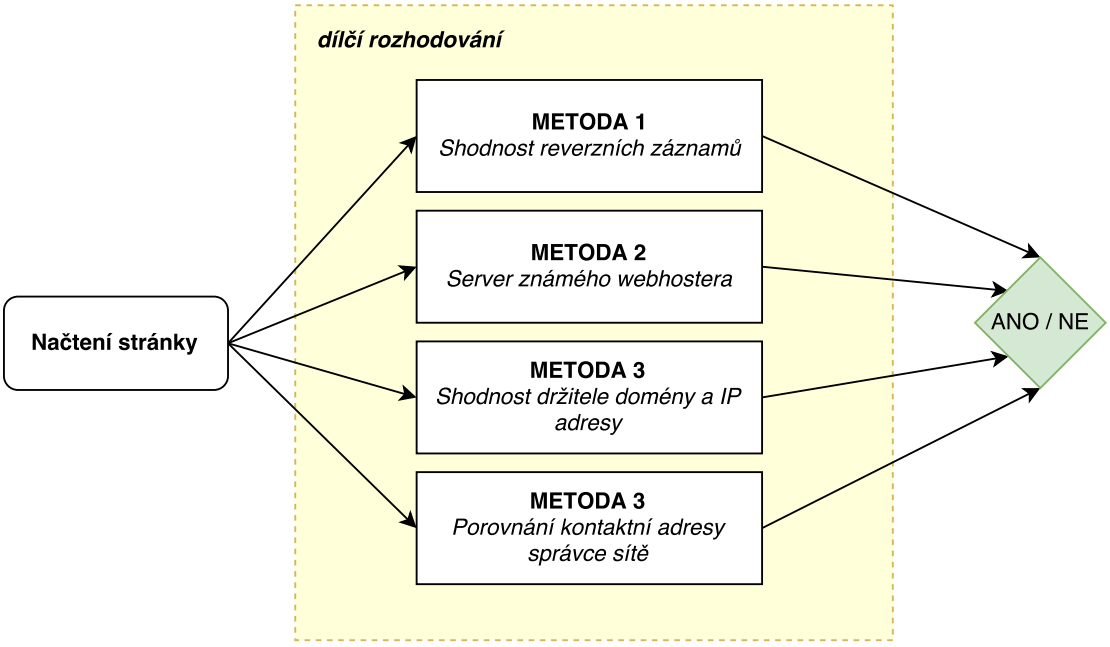
V případě, že je skóre kladné číslo včetně nuly, je považována stránka za hostovanou. Z tabulky 2.5 je patrné, že metody, které naznačují, že stránka není umístěna na sdíleném hostingu, skóre odečítají. Jediný případ, kdy je skóre zvýšeno, je v případě, že web je umístěn na serveru některého ze známých poskytovatelů hostingových služeb.

Pokud nedojde ke korektnímu vyhodnocení metody, není připočteno žádné skóre, tedy výsledek metody je vždy nulový. K nevyhodnocení metody může dojít, pokud se například nepodaří přeložit IP adresu serveru na doménové jméno, poté bude výsledek první metody uvažován jako nulový.

Jinými slovy, implicitně se očekává, že stránka je hostována a výzkumné metody musí vyhodnocovací algoritmus přesvědčit, že webová stránka hostována není.

Tab. 2.5: Vliv vyhodnocovacích metod na skóre určení hostování stránky na sdíleném hostingu.

Metoda	Ovlivnění skóre - Hostovaná stránka	
	Výsledek je pozitivní	Výsledek je negativní
1	0	-0,5
2	0,8	0
3	0	-0,4
4	0	-0,6



Obr. 2.1: Postup rozhodnutí detekované stránky.

3 REALIZOVANÁ APLIKACE PRO DETEKCI HOSTOVANÝCH STRÁNEK

Cílem realizační části této práce je vytvoření aplikace, která bude schopna detekovat hostování webových stránek na webhostingových službách. Z hlediska organizace vstupu a výstupu této aplikace je na jejím vstupu očekáván seznam webových stránek, resp. doménových jmen registrovaných v České republice. Realizovaná aplikace přijímá tento seznam ve formátu Comma-separed Values (CSV), tedy textovém souboru se sloupci oddělenými určeným oddělovačem. Výstupem aplikace je odhad počtu hostovaných stránek na sdílených webhostinzích pro zvolenou kategorii webů. V této kapitole práce je popsána architektura a základní algoritmus vytvořené aplikace.

3.1 Architektura aplikace

Realizovaná aplikace byla v souladu se zadáním práce vytvořena v interpretovaném programovacím jazyce Python 3 [18]. Pro zajištění možnosti spuštění aplikace nad velkým množstvím vstupních webových serverů byl zvolen přístup, kdy jsou data během vyhodnocení každého vstupního souboru z n celkových vstupních souborů průběžně ukládána do databáze. To umožňuje vyhodnocení v jeho průběhu bez ztráty integrity dat přerušit a pokračovat v něm dále později, případně do existující vyhodnocené kategorie vložit další zdrojové webové stránky.

Data, které jsou ukládána, jsou strukturovaná, a ačkoliv je výsledný databázový model, podrobněji popsaný v podkapitole 3.2, minimalistický, je mezi jeho entitami možné definovat vzájemné relační vztahy. Z tohoto důvodu byl pro aplikaci použit Systém řízení báze dat (SŘBD) [19]. Pro zajištění snadné přenositelnosti aplikace a její nekomplikované instalace, bylo zvoleno použití databázového stroje SQLite, který je integrován do základní knihovny jazyka Python [27] a prostřednictvím modulu `sqlite3` jazyka Python je programátorovi poskytováno rozhraní Application Programming Interface (API) pro práci s tímto integrovaným databázovým strojem. SQLite nepoužívá zvláštní aplikační server, který pracuje mezi klientskou knihovnou a souborem, ve kterém jsou uchována databázová data, proto není možné připojení k databázi jiným počítačem, a tedy využití této databáze je omezeno v případě, kdy by mělo být vyhodnocování webů distribuováno mezi více počítačů. Díky tomu je ale zajištěna funkčnost bez nutnosti konfigurace, podporou transakčních dotazů je možné využívat jeden databázový soubor více aplikacemi nebo jejími vlákny [20]. SQLite poskytuje plnohodnotnou implementaci strukturovaného dotazovacího jazyka Structured Query Language (SQL).

Implementace aplikace byla rozdělena do několika na sobě závislých modulů, přičemž moduly jsou využity k provádění dílčích částí vyhodnocování. Mnohé z těchto modulů umožňují jejich samostatné volání a přijímají argumenty příkazové řádky, podrobný popis těchto modulů a jejich přijímaných parametrů je uveden v příloze B. Veškeré moduly pracující s uživatelským vstupem využívají konzolové prostředí a pracují s argumenty příkazové řádky.

3.2 Databázový model aplikace

V rámci aplikace jsou veškerá získaná data určena k další analýze uchována v zabudované SQLite databázi `evaluation.db`. Z této databáze mohou být vypsány ať již celkové výsledky nebo dílčí výsledky pro jednotlivé kategorie. Současně jsou v této databázi uchovány informace o unikátních webových serverech, které obsluhovaly zaslané požadavky. Databáze obsahuje tři tabulky, jejichž struktura a význam je dále popsán.

Tabulka „results“

V tabulce `results` se nachází záznamy pro každý unikátní web, který byl vyhodnocen a výsledky jednotlivých detekčních metod. Současně jsou zde uloženy různé další informace získané z databází WHOIS, zaznamenané informace z hlavičky odpovědi, IP adresy webového serveru, který byl v době zpracování výsledkem DNS dotazu a také volitelnou poznámku ze vstupního souboru. Význam všech sloupců popisuje tabulka 3.1.

Tabulka „webhoster“

V tabulce `webhoster` se nachází seznam známých hostingů. Tento seznam je použit pro rozhodování o hostování pomocí druhé metody, viz kapitola 2.2. Implementovaný modul aplikace pro obsluhu databáze umožňuje přidávat záznamy do této tabulky (viz kapitola 3.3.5).

Výčet a význam datových položek databázové tabulky je uveden v následující tabulce 3.2.

Tabulka „webserver“

Do tabulky `webserver` jsou ukládány informace o každém webovém serveru, který byl výsledkem DNS dotazů během prohledávání webových serverů. Pokud je zpracována webová stránka, která je obsluhována serverem, který již byl uložen, je pro

Tab. 3.1: Datové sloupce tabulky „results“.

Název sloupce	Datový typ	Význam dat
domain	TEXT	Extrahovaná doména z URL adresy
timestamp	INTEGER	Datum a čas navštívení stránky
hosted_rev	NUMERIC	Výsledek metody č. 1 (shodný reverzní záznam)
hosted_known	NUMERIC	Výsledek metody č. 2 (známí webhosteři)
hosted_whois	NUMERIC	Výsledek metody č. 3 (shodnost WHOIS informací)
hosted_email	NUMERIC	Výsledek metody č. 4 (dle kontaktní adresy správce sítě)
server_httpd	TEXT	Prezentovaná signatura webového serveru
server_xpowerer	TEXT	Prezentovaná signatura aplikačního prostředí serveru
server_type	TEXT	MIME typ a kódování
server_ip	TEXT	Zjištěná IPv4 adresa serveru
server_ip_name	TEXT	Zjištěné doménové jméno IPv4 adresy serveru
server_ip_domain	TEXT	Extrahované doménové jméno IPv4 adresy serveru.
dnssec_keyset	TEXT	Použitá sada klíčů systému DNSSEC
domain_holdername	TEXT	Informace o držiteli domény z registru CZ.NIC
ip_holdername	TEXT	Informace o držiteli IP adresy z WHOIS registrů
category	TEXT	Kategorie, ze které byl web zařazen
original_url	TEXT	Původní URL požadavku před zpracováním.
note	TEXT	Volitelná poznámka z původního vstupního souboru.
https_support	NUMERIC	Informace o podpoře protokolu HTTPS (ano/ne)
ipv6_addr	TEXT	Zjištěná IPv6 adresa serveru

Tab. 3.2: Datové sloupce tabulky „webhoster“.

Název sloupce	Datový typ	Význam dat
hoster_domain	VARCHAR	Společností užívané doménové jméno
company	VARCHAR	Název společnosti

Tab. 3.3: Datové sloupce tabulky „webserver“.

Název sloupce	Datový typ	Význam dat
ip	VARCHAR	IPv4 adresa webserveru
timestamp	TIMESTAMP	Datum a čas získání informací
as	VARCHAR	Autonomní systém sítě
organization	VARCHAR	Organizace mající přidělený daný IP rozsah
isp	VARCHAR	ISP pro danou IP adresu
country	VARCHAR	Země umístění dané IP adresy
lat	NUMERIC	Zeměpisná šířka umístění dané IP adresy
lon	NUMERIC	Zeměpisná délka umístění dané IP adresy
city	INTEGER	Město, ve kterém je umístěná daná IP adresa
region	VARCHAR	Region, ve kterém je umístěna daná IP adresa
postal_code	VARCHAR	PSČ, ve kterém je umístěna daná IP adresa
contact_email	VARCHAR	Kontaktní mail správce pro danou IP adresu
contact_address	TEXT	Kontaktní poštovní adresa pro danou IP

urychlení běhu aplikace a snížení množství odesílaných požadavků namísto opětovného dotazování registrů a testování vlastností serveru vrácen výsledek právě z této tabulky. Získaná data o webových serverech jsou v práci použity pro obecný popis vlastností webových serverů v České republice, nacházející se v další, analytické, kapitole 4. Význam všech jejích datových sloupců je popsán v tabulce 3.3.

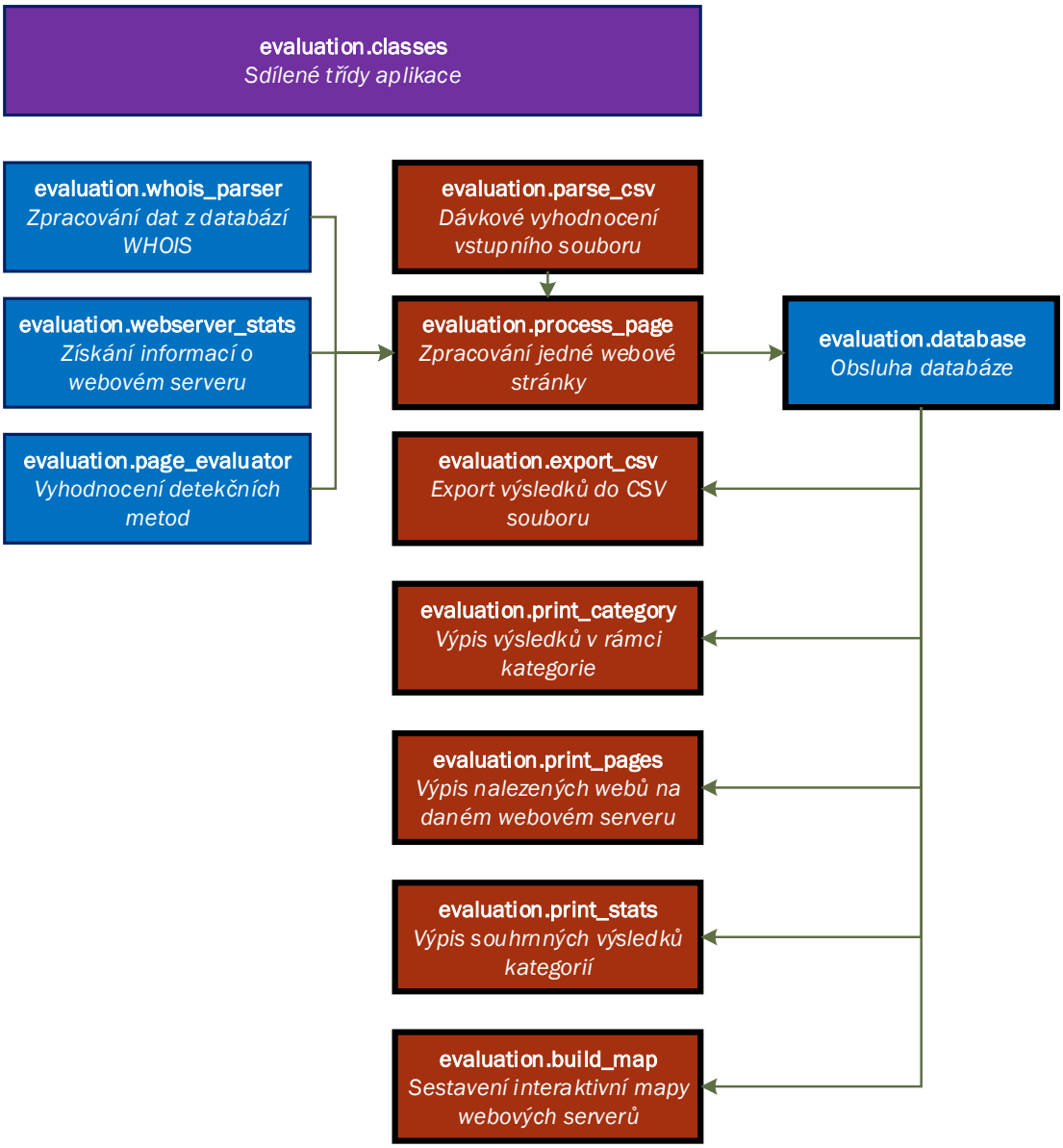
Tato tabulka je s tabulkou `results` vázána pomocí cizího klíče vazbou 1:N sloupci `ip` a sloupcem `server_ip` v tabulce `results`.

3.3 Popis modulů aplikace

Činnost aplikace je možné rozdělit do několika základních částí – zpracování jedné konkrétní stránky, dávkové zpracování jednoho vstupního souboru, prezentace dat a nástrojů pro automatizované vytvoření vstupního seznamu webových stránek. Kromě toho je implementována řada pomocných modulů pro poskytování dat ze vzdálených databází, práci s lokální SQLite databází, moduly pro práci s identifikátory URL. Hlavní moduly, jejich vzájemné závislosti a směr výměny informací mezi nimi je znázorněn na obrázku 3.1. Pomocné moduly jsou vyznačeny modrou barvou. Moduly, které jsou ohraničeny tučnou černou čarou přijímají argumenty z příkazové řádky.

3.3.1 Modul pro vyhodnocení webové stránky

Modul `process_page` je základním stavebním kamenem aplikace, který zajišťuje zpracování jedné konkrétní stránky a prostřednicím API pomocných modulů získává



Obr. 3.1: Architektura modulů realizované aplikace.

informace z externích zdrojů. Algoritmus vyhodnocení je zobrazen na obrázku 3.2.

Hlavní funkcí modulu je funkce `process_domain`, která na vstupu očekává zpracované doménové jméno ve formě objektu třídy `PreparedDomain` a kategorii webové stránky. Základní informace o průběhu jsou průběžně vypisovány na standardní výstup. Funkce může být zavolána prostřednictvím API modulu (z modulu `process_csv`) nebo pomocí parametrů předaných argumenty příkazové řádky. Funkce ve výchozím nastavení přeskakuje již jednou vyhodnocené stránky (chování lze upravit nastavením parametru `reevaluate`), v případě nastavení argumentu `restoreconn` je požadováno po externích modulech zavolání uživatelského skriptu `restoreconn.sh` za účelem obnovení spojení při detekci výpadku k externím registrům. Výchozím chováním, viz obr. 3.2, je přeskakování domén, které nejsou v „.cz“ zóně, případně jsou doménami třetí a vyšší úrovně (lze přenastavit nastavením parametrů `cz_ommit` a `subdom_ommit`). Pokud je funkci předán parametrem `output` objekt pro předání informací o vyhodnocení, jsou zásadní informace o výsledku uloženy do tohoto objektu.

3.3.2 Modul pro zpracování vstupního souboru

Modul `process_csv` zajišťuje načtení vstupního souboru z textového formátu oddělovaného definovaným oddělovačem, jako výchozí se používá znak tabelátoru.

Hlavní funkce modulu `parse_file` načte zadaný textový soubor, pro každý jeho řádek rozdělí a ze zadaného sloupce extrahuje URL jednotlivých webových stránek v tomto vstupním souboru. Pro každou doménu zavolá funkci `process_domain` modulu `process_page`. Informace z této funkce uchovává do objektu `output` a pro každou načtenou stránky jsou tyto informace následně uloženy do souboru `output.txt`.

Parametry této funkce jsou číslo sloupce s webovými adresami, oddělovače sloupců, počet řádků vstupního souboru, které se považují jako hlavička souboru, kategorii vstupního souboru a parametry, které se předávají funkci `process_page`, jako je nastavení přeskakovaných stránek.

Pomocí parametru `notecolumn` může být definován sloupec s poznámkou ze vstupního souboru, který je uložen do databáze SQLite. V případě, že je definován parametr `noterow`, je celý vstupní řádek do databáze uložen ve formátu JSON. Modul přijímá své parametry prostřednictvím argumentů příkazové řádky, jejich popis je uveden v příloze B.

3.3.3 Modul pro načtení informací z databáze WHOIS

Modulem `whois_parser` je zajištěno zpracování dat získaných z registru CZ.NIC a registrů jednotlivých registrů přidělených IP adres. Dále modul načítá informace

o dané IP adrese serveru z geolokační databáze GeoIP s využitím REST API rozhraní služby IP-API.com.

Implementované funkce tohoto modulu tvoří funkce `probe_domain` pro zjištění informací o doméně z databáze CZ.NIC, `probe_contact` pro stažení kontaktu z databáze CZ.NIC, funkce `probe_ip` pro zjištění informací o IP adrese serveru z WHOIS databáze a k ní příslušným geolokačních údajům.

Funkce mohou být volány prostřednictvím API modulu nebo argumentů příkazové řádky. Výsledky jsou modulem vráceny jako asociativní pole. Většina funkcionality je implementována ve třídách `CZNICWhoisResult`, `IPWhoisResult` a `GeoIPResult`. Z abstraktní třídy `CZNICWhoisResult` jsou odvozeny třídy `DomainWhoisResult` a `ContactWhoisResult`.

Funkcím může být definován parametr `restore_conn`. V případě, kdy je nastaven a je detekováno opakované selhání odpovědi z registru, je zavolán externí skript `restore_conn.sh`, kterým může být problém automatizovaně vyřešen. Až po několika selhání je běh skriptu přerušen.

3.3.4 Modul pro zjištění informací o webovém serveru

Modul `webserver_stat` zajišťuje sekvenci úlohy, kterými je popsáno chování webového serveru, zejména pak hlaviček, které byly součástí přijaté odpovědi od HTTP serveru.

Hlavní funkce `probe_server` přijímá jako svůj parametr adresu webu k přijetí odpovědi, pomocí parametru `force_https` lze vynutit využití HTTPS připojení, čehož se využívá při testování funkčnosti HTTPS protokolu na daném webu. Funkce umístěná v tomto modulu se zaměřuje na navázání HTTP spojení s cílovým serverem. Vzhledem k možnosti virtuálních webů na webovém serveru je uchovávána informace o hlavičce zjišťována pro každý web, protože pro každý web na daném serveru se mohou sledované informace z hlavičky lišit, viz kapitola 1.7. Z tohoto důvodu jsou tyto informace ukládány přímo do tabulky `result`, dle kapitoly 3.2. Širší princip algoritmu zjištění informací o webovém serveru je naznačen na obrázku č. 3.3.

3.3.5 Modul pro obsluhu databáze

Modul `database` zajišťuje připojení k SQLite databázi a provádění dotazů nad otevřenou databází. Obsahuje dvě globální proměnné `con` a `cur` uchovávající kontext připojení k databázi a kurzor databázového klienta interpretu Python.

Hlavními funkcemi pro manipulaci s daty jsou funkce `exec_query` sloužící k provedení jednoho SQL dotazu a funkce `commit` zajišťující potvrzení transakce a za-

psání změn do databáze. Připojení a odpojení k databázi je realizováno funkcemi `open_connection` a `close_connection`. Tyto funkce jsou volány závislými komponentami prostřednictvím API modulu.

Dále jsou definovány funkce `run_script` a `truncate_table` pro provádění víceřádkových operací. Pro inicializaci výchozího databázového schématu je připravena funkce `init_db`.

Tento modul umožňuje volat několik operací pomocí argumentů příkazové řádky, tyto operace jsou popsány v příloze B.

Schéma SQLite databáze

Při volání funkce `init_db` je vytvořena nová prázdná databáze, nad kterou je spuštěna následující SQL věta, zajišťující inicializaci databázového schematu dle podkapitoly 3.2. Tato SQL věta je vypsána ve výpisu v části přílohy B.2.

3.3.6 Moduly pro výpis výsledků z databáze

Moduly `print_category` a `print_stats` slouží pro výpis výsledků vyhodnocení stránek z SQLite databáze.

Modul `print_category` zajišťuje výpis výsledků v rámci jedné kategorie, modul `print_stat` vypisuje výpis výsledků jednotlivých kategorií. Výsledky jsou vypsány na standardní výstup ve formě textových tabulek.

V případě, že je nutné s daty dále pracovat, je vhodné použít modul `export_csv`, který zajistí export výsledků zvolené kategorie do CSV souboru.

Veškeré tyto moduly přijímají argumenty z příkazové řádky. Celkové určení, zda je stránka hostována na základě dílčích metod, je provedeno až při výpisu, aby bylo možné parametry rozhodování (viz kapitola 2.5) změnit bez nutnosti provádět znovu celé dotazování.

3.3.7 Modul pro výpis webů hostovaných na serverech

Modul `print_pages` poskytuje možnost výpisu všech webových stránek, které byly detekovány na konkrétním webovém serveru (zadaném jeho IPv4 adresou) nebo zadané adresy sítě (pomocí adresy sítě a masky sítě). Tento výpis je realizován na základě nalezené shody IP adresy serveru u více webů.

Tento modul přijímá vstupní parametry z příkazové řádky. Pomocí přepínače `webhosting` lze definovat doménové jméno konkrétního webhostera definovaného v databázové tabulce `webhoster`. V případě jeho aktivace jsou vypsány webové stránky, které byly detekovány na webových serverech daného známého poskytovatele webhostingu. Pokud je navíc při volání modulu uveden parametr `summary`, je

zobrazen souhrn počtu detekovaných webových stránek umístěných u daného webhostera a také minimální, maximální a průměrný počet umístěných webů na jednom serveru.

Relevantní výsledky této metody je možné očekávat pouze s dostatečně velkou množinou vyhodnocených webů v databázi.

3.3.8 Modul pro export dat z databáze

Modulem `export_csv` poskytuje možnost exportovaných výsledků z SQLite databáze do výstupního souboru formátu CSV. Funkcionalita pro export dat je implementována třídě `EvaluatedCategory` metodou `to_csv`.

Modul přijímá argumenty z příkazové řádky, implicitně jsou výsledky uloženy do souboru s názvem dle schématu `<kategorie>_export.csv`, ve kterém jsou jednotlivé sloupce odděleny tabelátorem.

3.3.9 Modul pro vizualizaci umístění webových serverů

Pomocí modulu `build_map` jsou informace o detekované poloze jednotlivých nalezených serverů zobrazeny do mapy. Generování mapy zajišťuje balíček `folium`. Výstupní soubory jsou uloženy ve formátu hypertextového dokumentu, ve kterém je připojen skript, který zajistí vykreslení definovaných entit do mapy ve webovém prohlížeči pomocí jazyku Javascript.

3.3.10 Definice tříd použitých v modulech

Součástí jmenného prostoru aplikace je soubor `classes.py`, ve kterém je definován předpis, pro několik tříd, které jsou použity v aplikaci. Motivací pro oddělení těchto tříd do souboru je snaha o zvýšení znovupoužitelnosti a přehlednosti implementovaného zdrojového kódu. V aplikaci jsou definovány níže popsané třídy.

- `ContactWhoisResult` je třída, která uchovává informace o výsledky dotazu na entitu kontaktu načtenou z registru databáze `CZ.NIC`. Třída je odvozená z mateřské třídy `CZNICWhoisResult`. Její chování doplňuje o definici metody `process_query` pro provedení dotazu specifického dotazu na konkrétní kontakt v registru.
- `CZNICWhoisResult` je abstraktní třída, která definuje základní funkcionalitu práce s registrem CZ.NIC. Při vytvoření objektu je konstruktorem provedeno spuštění dotazu na databázi metodou `process_query`.

- `DomainWhoisResult` je třída, která uchovává informace o výsledky dotazu na entitu doménového jména načtenou z registru databáze `CZ.NIC`. Třída je odvozená z mateřské třídy `CZNICWhoisResult`. Její chování doplňuje o definici metody `process_query` pro provedení dotazu na konkrétní doménové jméno.
- `EvaluatedCategory` je třídou, ve které jsou uloženy informace o celkovém výsledku vyhodnocené kategorie. Ve své třídní proměnné `hosted_page` jsou uloženy nalezené detekované hostované weby v dané kategorii. Vestavěná metoda pro výpis třídy vrátí tabulku se seznamem webů v kategorii a jim příslušným výsledkům. Na závěr výpisu je připojena tabulka obsahující statistiku dané kategorie. V této třídě je také definována metoda pro export výsledků do souboru `to_csv`.
- `EvaluatedPage` je třída, která uchovává informace o výsledku jednotlivých detekčních metod pro jeden konkrétní web. Při vytvoření objektu je zavolána metoda `calc_score`, která na základě parametrů předaných třídním konstruktorem provede výpočet celkového skóre vyhodnoceného webu, viz kapitola 2.5.
- `GeoCoordinates` je třída, ve které je uchován pár souřadnic zeměpisné délky a šířky.
- `GeoIPResult` je třída, ve které je uchován výsledek dotazu na geolokační službu. Po vytvoření je konstruktorem zavolána metoda `get_from_geoip`, ve které jsou pomocí knihovny `urllib` staženy a dekódovány informace z geolokační služby GEO-IP.
- `IPWhoisResult` je třída, která ve svých členských proměnných uchovává výsledek dotazu na WHOIS databázi pro příslušnou IP adresu.
- `PreparedDomain` je pomocná třída, ve které uložena trojice informací vztahující se k doméně, která je vstupem pro detekční algoritmus. Jde o extrahované doménové jméno, originální URL požadavku a případnou poznámku přečtenou ve zdrojovém souboru.

Vzájemné relace mezi těmito třídami a jejich členské proměnné a dostupné metody zobrazuje UML diagram na obr. 3.4.

3.3.11 Skripty pro generování vstupních souborů

K aplikaci jsou přiloženy dva nezávislé skripty, `build_db_firmy` a `build_db_toplist`. Tyto skripty umožňují stažení seznamu webových stránek umístěných v zadané kategorii z portálů Firmy.cz a Toplist.cz.

Skripty zpracovávají odpověď webového serveru těchto katalogů a extrahují z této odpovědi domény webů nalezených ve zvolené kategorii. Vzhledem k tomu, že výsledky obou těchto katalogů jsou stránkované, hlavní funkce `collect_category` je rekurzivně volána při nalezení odkazu na další stránku výpisu.

Skript přijímá řetězec reprezentující část URL požadované kategorie těchto katalogů a cestu k výstupnímu souboru. Pomocí volitelného argumentu lze specifikovat číslo stránky, na kterém má být zpracování zahájeno. To je užitečné v případě nutnosti navázat na přerušený běh skriptu.

3.4 Výpočetní složitost detekčního algoritmu

Vyhodnocení jedné webové stránky je realizováno spuštěním sekvencí úloh, které musí být před zápisem do databáze provedeny. Je provedeno několik dotazů na WHOIS databáze, jednak na databázi sdružení CZ.NIC, dále na databáze organizací provozující registry přidělených IP adres. Následuje zaslání požadavku DNS resolveru pro na přeložení doménového jména na IPv4 a IPv6 adresu, posléze je pomocí dalších DNS dotazů přeložena na doménové jméno. Webová stránka je také navštívena pomocí protokolu HTTP i HTTPS a čeká se na odpověď po dobu stanoveného časového limitu. Na době zpracování se podílí také odezva mezi testovací aplikací a cílovými servery.

Díky práci s tolika vzdálenými externími zdroji je rozdílnost doby zpracovávání jednotlivých webů velká a nelze určit čas, jak dlouho trvá vyhodnocení jedné stránky. Z hlediska závislosti doby zpracování na množství vstupních webů je časová složitost detekčního algoritmu lineární (n) [21].

Vzhledem k tomu, že jsou výsledky průběžně ukládány do relační databáze, jsou po zpracování stránky pracovní data vztahující se k této stránce uvolněny z operační paměti. Prostorová složitost algoritmu je tedy konstantní (c) [21].

Průměrná doba zpracování

Průměrná doba zpracování byla zjištěna experimentem, kdy byl po dobu zpracování jedné dávky vstupních dat měřen čas mezi provedením prvního a posledního záznamu. Aplikace byla spuštěna na serveru umístěném v brněnském datovém centru a na tomto serveru nebyly v čase experimentu spuštěny žádné náročné procesy. Výsledky experimentu jsou uvedeny v tabulce 3.4. Vstupy, které byly v úvodu běhu skriptu přeskočeny, jsou ve výpočtu průměrné doby zpracování webu zanedbány, protože uplynulý čas mezi spuštěním algoritmu a odmítnutím takových vstupů je vzhledem k době zpracování celého záznamu zcela zanedbatelný.

Tab. 3.4: Výsledky experimentálního zjištění průměrné doby zpracování.

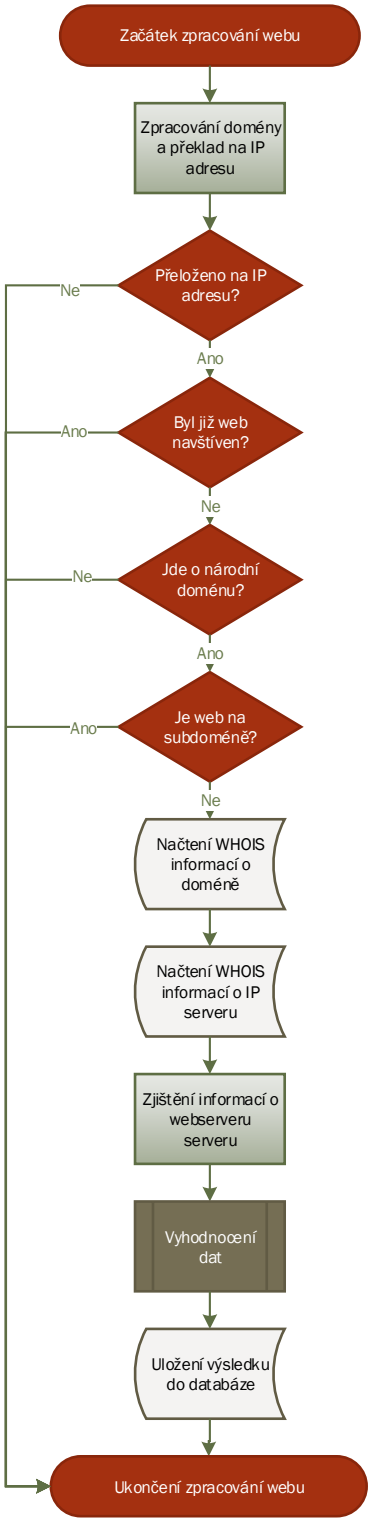
Celkový počet vstupů:	2 200
Počet odmítnutých vstupů:	540
Počet vyhodnocených stránek:	1 660
Celková doba běhu skriptu:	45 m 10 s
Průměrná doba zpracování webu:	1,63 s

3.5 Příklad vyhodnocení

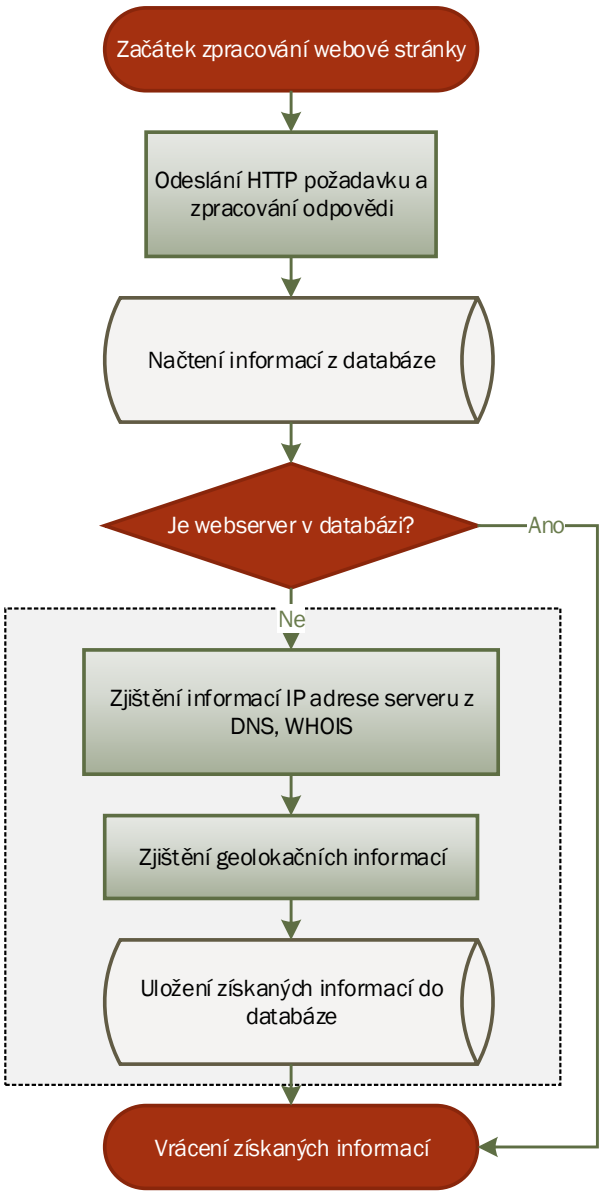
Pro ilustraci činnosti detekčního algoritmu následuje vzorové vyhodnocení všech detekčních metod a určení celkového skóre dle popsané metodiky. Příklad je proveden na webové stránce Vysokého učení technického v Brně, jehož webová stránka je `www.vutbr.cz`. Pro vyhodnocení jsou provedeny následující kroky, které se přímo podílí na výsledku detekce.

1. Doménové jméno je rozděleno podle doménových úrovní a je ověřeno, že se jedná o národní .cz doménu.
2. Dále je ověřeno, že web není na subdoméně (kromě subdomény `www`).
3. Web je přeložen rekurzivním DNS dotazem na IP adresu `147.229.2.90`.
4. IP adresa je přeložena na doménové jméno `piranha.ro.vutbr.cz`.
5. Jsou načteny informace z registru WHOIS pro doménu `vutbr.cz`. Uvedená organizace v kontaktu pro doménu je „Vysoké učení technické v Brně“.
6. Jsou načteny informace z registru WHOIS pro IP adresu webového serveru `47.229.2.90`. Uvedená organizace v kontaktu registru je „Brno University of Technology“.
7. Z výsledku překladu IP adresy serveru na doménové jméno je zjištěno a její extrakce, že doména 2. úrovně `vutbr.cz` webu a doménového jména IP adresy serveru je shodná, první metoda je tedy vyhodnocena jako **nesplněná** (stránka není hostována).
8. Z výsledku překladu IP adresy serveru na doménové jméno je zjištěno, že doména 2. úrovně `vutbr.cz` není na seznamu známých webhostingů, druhá metoda je tedy vyhodnocena jako **nesplněná** (stránka není hostována).
9. Je zjištěno, že jméno organizace, které je přidělena IP adresa je „Brno University of Technology“, avšak v registru je u domény uvedena organizace „Vysoké učení technické v Brně“ – tyto názvy nejsou vyhodnoceny jako podobné a třetí metoda je tedy vyhodnocena jako **splněná** (stránka je hostována).
10. Kontaktní e-mail uvedený u organizace, která má přidělenou příslušnou IP adresu serveru je umístěn v rámci dotazované domény, čtvrtá metoda je tedy vyhodnocena jako **nesplněná** (stránka není hostována).
11. Na základě hodnot uvedených v tabulce 2.5 je vypočteno celkové skóre:
    $Skóre = (-0,5) + 0 + 0 + (-0,6) = 1,1.$
    Výsledkem je záporné číslo a stránka je tedy celkově vyhodnocena jako **nehostovaná na sdíleném webhostingu**.

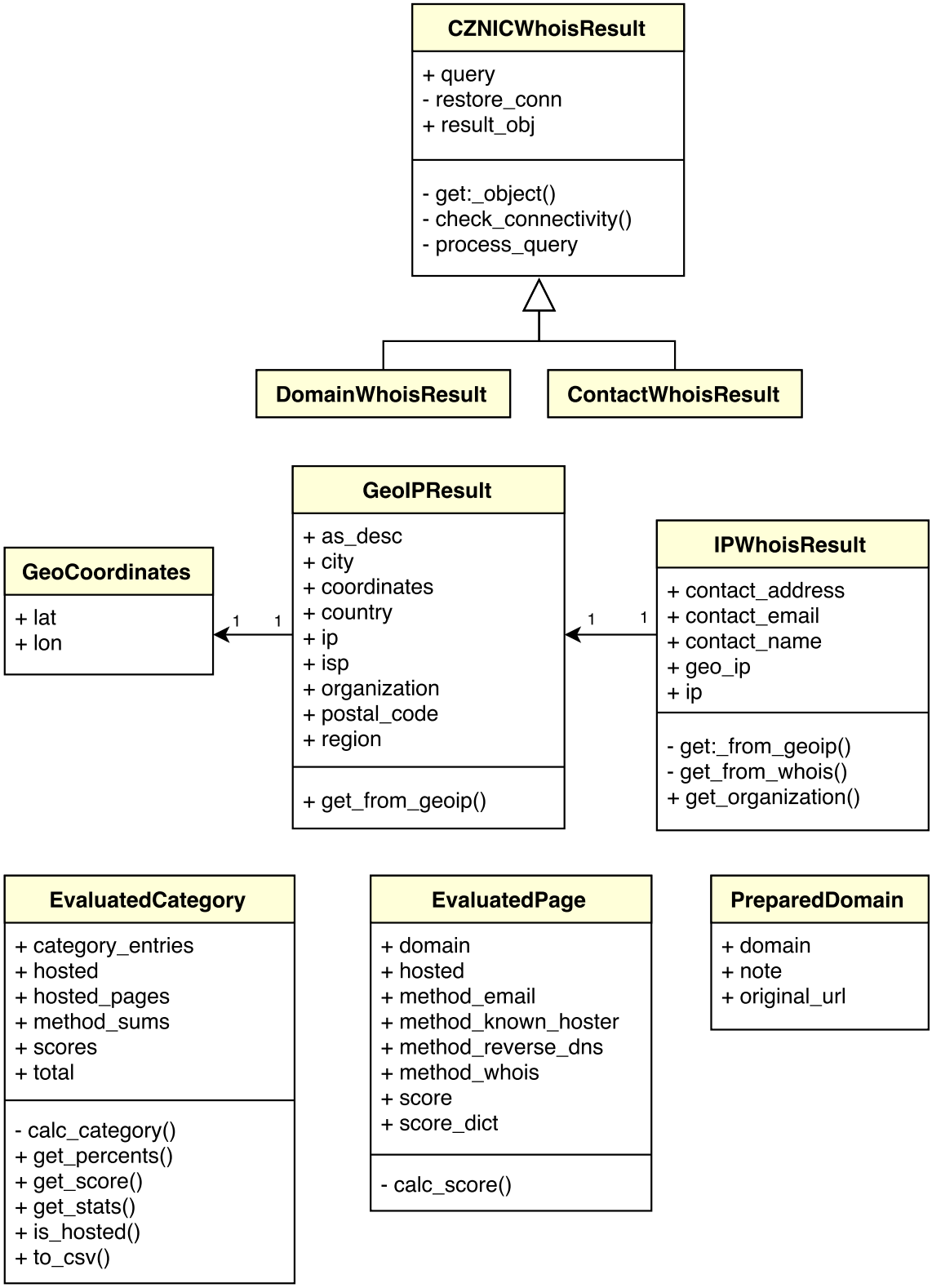
Z postupu algoritmu je patrné, že třetí metoda nebyla vyhodnocena správně díky rozdílným použitým jazykům organizace, avšak díky vlivu ostatních metod je konečný výsledek správný.



Obr. 3.2: Vývojový diagram vyhodnocení jedné webové stránky. Popis vyhodnocení získaných dat (reprezentovaný blokem „Vyhodnocení dat“) je popsán v kapitole 2.5.



Obr. 3.3: Vývojový diagram zjištění informací o webovém serveru.



Obr. 3.4: UML diagram tříd aplikace.

4 ANALÝZA ZÍSKANÝCH DAT

V následující kapitole je provedena analýza situace ohledně hostování webových stránek v České republice. Pro provedení této analýzy byly využity dříve popsané vytvořené nástroje, které byly spuštěny na několika vstupních sadách webových stránek. Každá tato sada reprezentuje určitou skupinu subjektů provozovatelů daných webových stránek. Díky tomuto rozdělení je možné sledovat odchylky výsledků analýzy v těchto skupinách.

Vzhledem k popsané metodice rozhodování hostování webových stránek jsou zpracovány pouze webové prezentace přístupné na národní doméně druhého řádu, tedy např. subdomény typu „organizace.brno.cz“ jsou ve zpracování přeskočeny.

4.1 Zkoumané skupiny subjektů

Pro analýzu byly zvoleny skupiny subjektů provozovatelů webových stránek dle tabulky. V tabulce je uveden počet získaných unikátních webových domén v dané kategorii. Pokud není uvedeno jinak, domény byly získány z katalogů `http://www.firmy.cz` a `http://www.toplist.cz`. Veškerá sesbíraná data byla použita pouze za účelem provedení popsané analýzy.

Tab. 4.1: Analyzované skupiny provozovatelů webů.

Kategorie subjektů	**Počet získaných domén**
Vysoké školy [1]	221
Střední školy	1 305
Nemocnice	536
Internetové obchody	13 131
Řemeslníci	8 604
Poskytovatelé webhostingových služeb [2]	145
Státní a vládní organizace	124
Banky	45
Komerční pojišťovny	31
Realitní kanceláře [3]	2 431
Celkem	26 573

[1] Zpracováno dle [22].
[2] Vlastní zpracování.
[3] Zpracováno dle [23].

Takto získané seznamy byly použity jako vstupní data do aplikace. Celkem jde tedy o 26 573 webů v deseti kategoriích. Vzhledem k tomu, že ze seznamů nebyly odstraněny stránky, které jsou mimo českou doménu, jsou na subdoméně nebo daná webová stránka ať již z důvodu expirace domény, případně jiného důvodu nefunkčnosti webu v době prohledání testovacím skriptem nebyla funkční, je počet reálně korektně vyhodnocených webů nižší.

Vzhledem k tomu, že prohledávací skripty byly spouštěny pro jednotlivé kategorie opakovaně, je malá pravděpodobnost, že by některý z webů měl ve všechny sledované časy náhodný výpadek, a proto se dá očekávat, že takové stránky nejsou funkční.

Během procházení těchto webů obsluhovalo požadavky aplikace celkem 5 890 unikátních webových serverů.

4.2 Podíl hostovaných webových stránek

Spouštěním detekčních skriptů a vyhodnocení podle popsaných navržených metodik bylo aplikací realizovanou v rámci této práce detekováno hostování webů na jednotlivých webech organizovaných dle kategorií subjektů provozovatelů webů, viz kapitola 4.1.

Podrobné výsledky jednotlivých metod a celkovém rozhodnutí, zda je stránka hostována na sdíleném webhostingu, jsou uvedeny v tabulce 4.2, ve které jsou uvedeny absolutní počty webů, pro které byly konkrétní detekční metody vyhodnoceny jako splněné a dále pak v tabulce 4.3, která představuje procentuální vyjádření podílu hostovaných stránek pro jednotlivé kategorie. Grafické znázornění výsledků je zobrazeno na obrázku 4.1.

Tab. 4.2: Počet hostovaných webů v jednotlivých kategoriích.

Kategorie	**Vyhod.**	**Počet hostovaných webů**				
		Met. 1	Met. 2	Met. 3	Met. 4	Výsledek
Banky	36	19	2	31	28	27
Eshopy	11 314	10 391	1 484	11 300	11 228	11 104
Nemocnice	469	419	89	467	452	450
Pojišťovny	26	20	1	25	22	22
Realitní kanceláře	1 660	1 515	270	1 654	1 628	1 629
Řemeslníci	6 563	6 134	1 686	6 445	6 510	6 402
Státni organizace	109	71	8	104	98	83
Střední školy	1 192	1 008	204	1 188	1 154	1 073
Vysoké školy	103	62	14	93	73	66
Webhostingy	39	11	15	29	24	11
Celkově	21 511	19 650	3 773	21 336	21 217	20 867

Tab. 4.3: Podíl hostovaných webů v jednotlivých kategoriích.

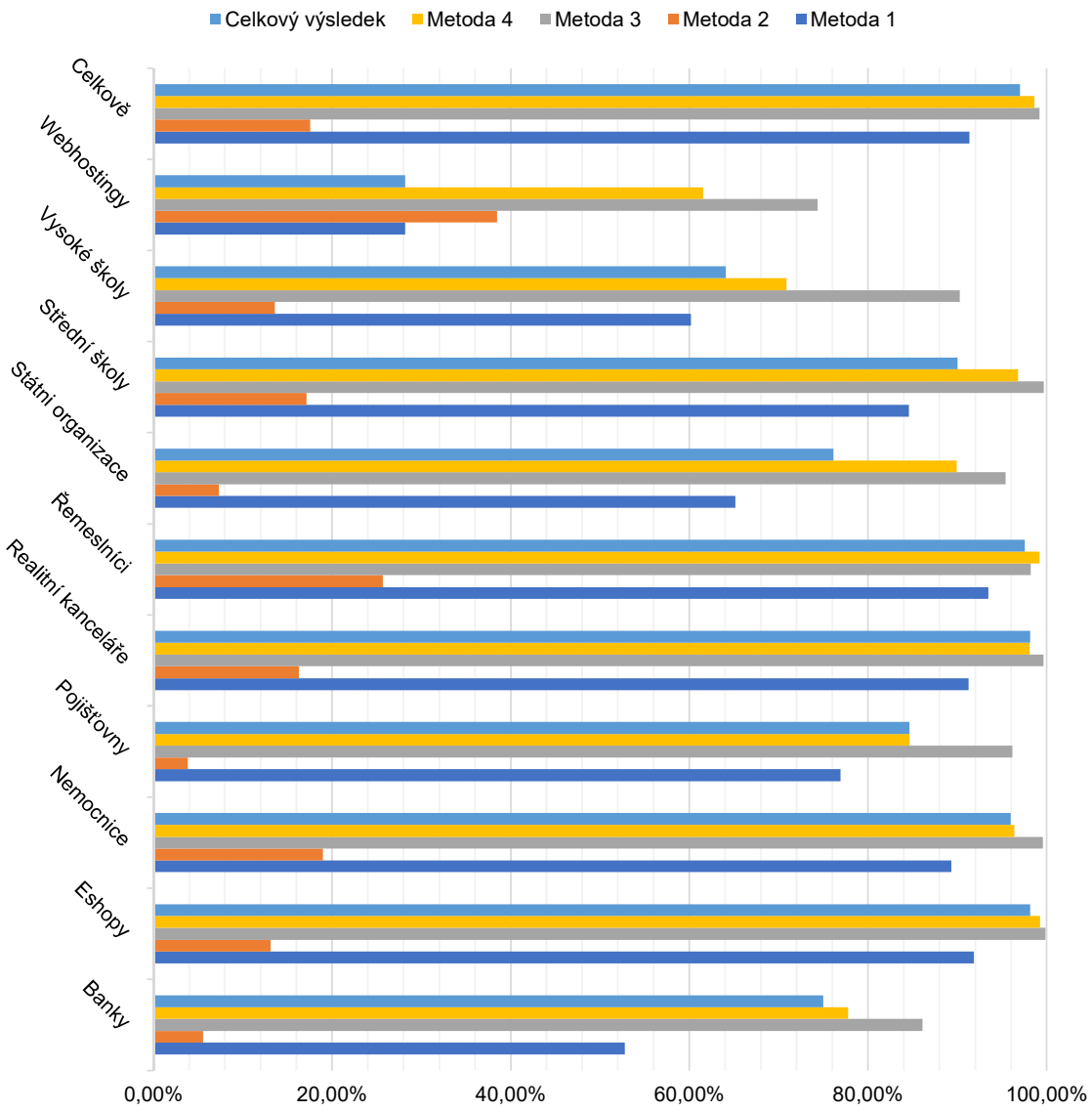
Kategorie	**Vyhodnocených**	**Podíl hostovaných webů**				
		Met. 1	Met. 2	Met. 3	Met. 4	Výsledek
Banky	36	52,78%	5,56%	86,11%	77,78%	75,00%
Eshopy	11 314	91,84%	13,12%	99,88%	99,24%	98,14%
Nemocnice	469	89,34%	18,98%	99,57%	96,38%	95,95%
Pojišťovny	26	76,92%	3,85%	96,15%	84,62%	84,62%
Realitní kanceláře	1 660	91,27%	16,27%	99,64%	98,07%	98,13%
Řemeslníci	6 563	93,46%	25,69%	98,20%	99,19%	97,55%
Státni organizace	109	65,14%	7,34%	95,41%	89,91%	76,15%
Střední školy	1 192	84,56%	17,11%	99,66%	96,81%	90,02%
Vysoké školy	103	60,19%	13,59%	90,29%	70,87%	64,08%
Webhostingy	39	28,21%	38,46%	74,36%	61,54%	28,21%
Celkově	21 511	91,35%	17,54%	99,19%	98,63%	97,01%

Z výsledků je zřejmé, že naprostá většina zkoumaných webů na české národní doméně je provozována na sdíleném webhostingu. Zároveň je nutné brát zřetel na fakt, že dílčí hodnoty jednotlivých metod jsou pouze informační, vzhledem k aditivnímu charakteru určení celkového výsledku je směrodatný pouze celkový výsledek všech metod uvedený ve sloupci „Výsledek“.

Ve zkoumaných kategoriích se nachází některé, ve kterých bylo očekáváno nadprůměrné množství webů hostovaných organizací – zejména šlo o kategorie bank, pojišťoven, státních organizací a vysokých škol. Zejména v grafickém vyjádření je toto patrné. Zejména v kategorii bank bylo překvapením, že jednotlivé subjekty mnohdy neměly web umístěný ve svém adresním rozsahu, DNS záznamy pro IP adresy jejich webových serverů ve většině případech ukazovaly na ISP, nikoliv přímo na danou organizaci, proto je zřejmé, že mezi expertním a strojovým posouzením umístění jejich webů by byl zřetelný rozdíl.

Největší množství serverů provozovaných samotnými institucemi je mezi vysokými školami, pokud není do tohoto porovnání uvažována kategorie webhostingů – v této kategorii není výsledkem nulový podíl hostovaných webů z podobných důvodů jako u bank, avšak v menším měřítku. Některé webhostingy provozují své služby pod více značkami, to se projevilo na zvýšení nepřesnosti všech detekčních metod.

Největší množství hostovaných webů bylo zaznamenáno v kategoriích internetových obchodů, webových prezentací řemeslníků a realitních kanceláří. V těchto kategoriích se pohybuje mnohem více menších subjektů, které provoz svých webů svěřují poskytovatelům webhostingových služeb.



Obr. 4.1: Přehled výsledků jednotlivých detekčních metod v rámci kategorií.

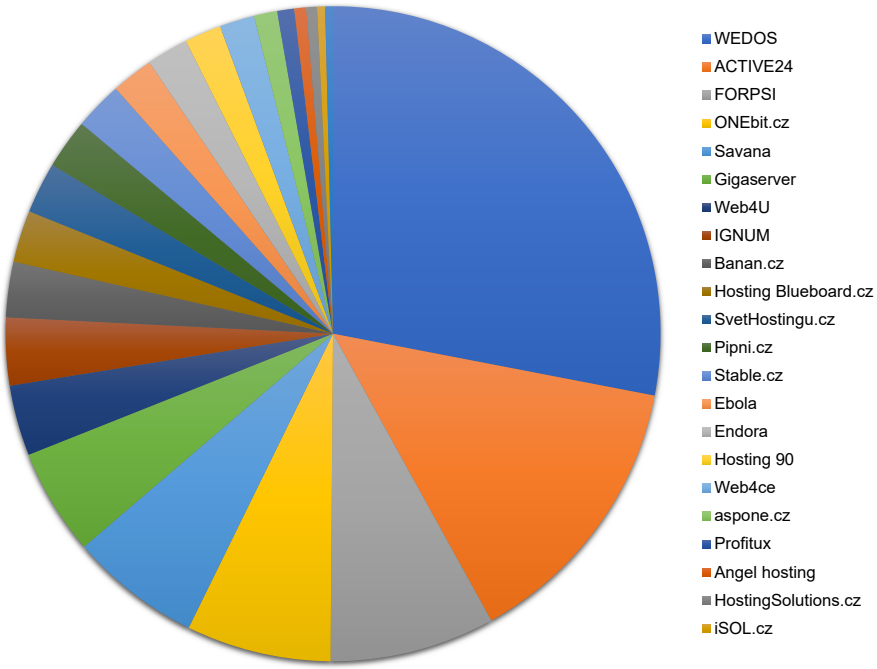
4.3 Zastoupení poskytovatelů webhostingu

Na základě porovnávání výsledků reverzních DNS dotazů pro IP adresy webových serverů bylo pro sledovaný vzorek (tj. 21 511 webů) určeno hostování webů u jednotlivých známých poskytovatelů hostingových služeb (uvedených v databázové tabulce `webhoster`). Výsledky této analýzy jsou uvedeny v tabulce 4.4. Podíl jednotlivých společností na hostování sledovaného vzorku je ilustrován také grafem 4.2. V tomto grafu jsou pro vyšší přehlednost jsou uvedeny pouze hostingy, které se podařilo určit a provozovaly alespoň 20 webů ze sledovaného vstupního vzorku. Ty weby,

které nebyly určeny (tj. asi 73 %), jsou buď hostované ve vlastní infrastruktuře nebo u menšího poskytovatele hostingu.

Tab. 4.4: Počet domén hostovaných u jednotlivých webhosterů

Provozovatel webhostingu	**Počet webů**	**Podíl**
Neurčeno	15767	73,3%
WEDOS	1610	7,5%
ACTIVE24	801	3,7%
FORPSI	468	2,2%
ONEbit.cz	410	1,9%
Savana	372	1,7%
Gigaserver	301	1,4%
Web4U	200	0,9%
IGNUM	192	0,9%
Banan.cz	157	0,7%
Hosting Blueboard.cz	147	0,7%
SvetHostingu.cz	145	0,7%
Pipni.cz	141	0,7%
Stable.cz	137	0,6%
Ebola	119	0,6%
Endora	118	0,5%
Hosting 90	104	0,5%
Web4ce	99	0,5%
aspone.cz	66	0,3%
Profitux	48	0,2%
Angel hosting	34	0,2%
HostingSolutions.cz	30	0,1%
iSOL.cz	23	0,1%
Thosting	22	0,1%



Obr. 4.2: Podíl zjištěných poskytovatelů webhostingových služeb.

4.4 Umístění serverů v autonomních systémech

Pro webové stránky byl zjištěn autonomní systém, ve kterém se web nachází. Protože velcí poskytovatelé webhostingu mají přidělené své autonomní systémy, je možné sledovat souvislost těchto údajů se zjištěným zastoupením poskytovatelů webhostingu ve zkoumaném vzorku v předchozí podkapitole.

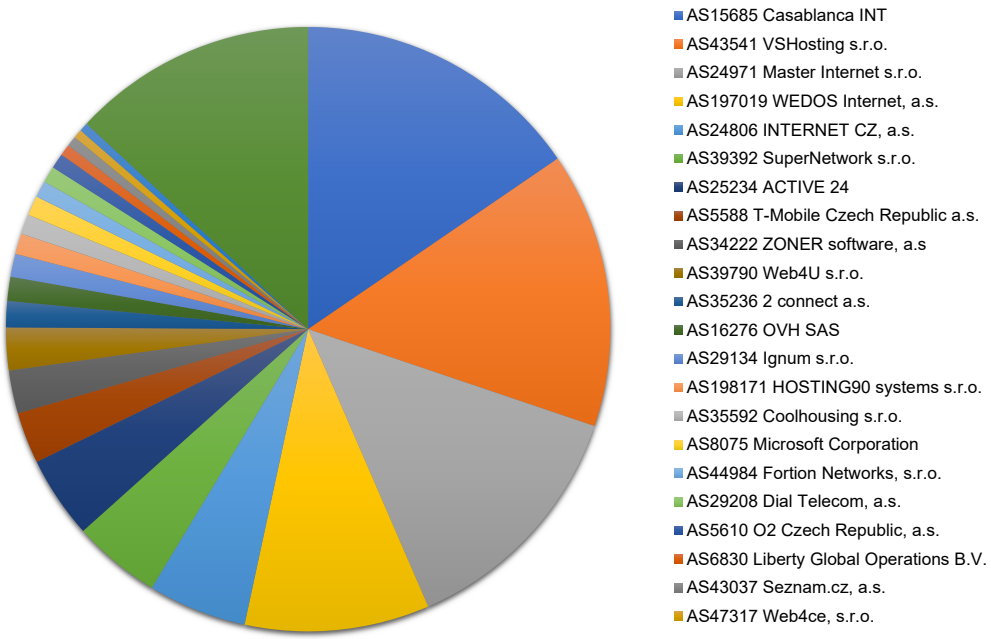
Seznam nejčastějších detekovaných autonomních systémů je uveden v tabulce 4.5. Zastoupení jednotlivých autonomních systémů pak ilustruje graf 4.3.

Tab. 4.5: Nejčastěji detekované autonomní systémy.

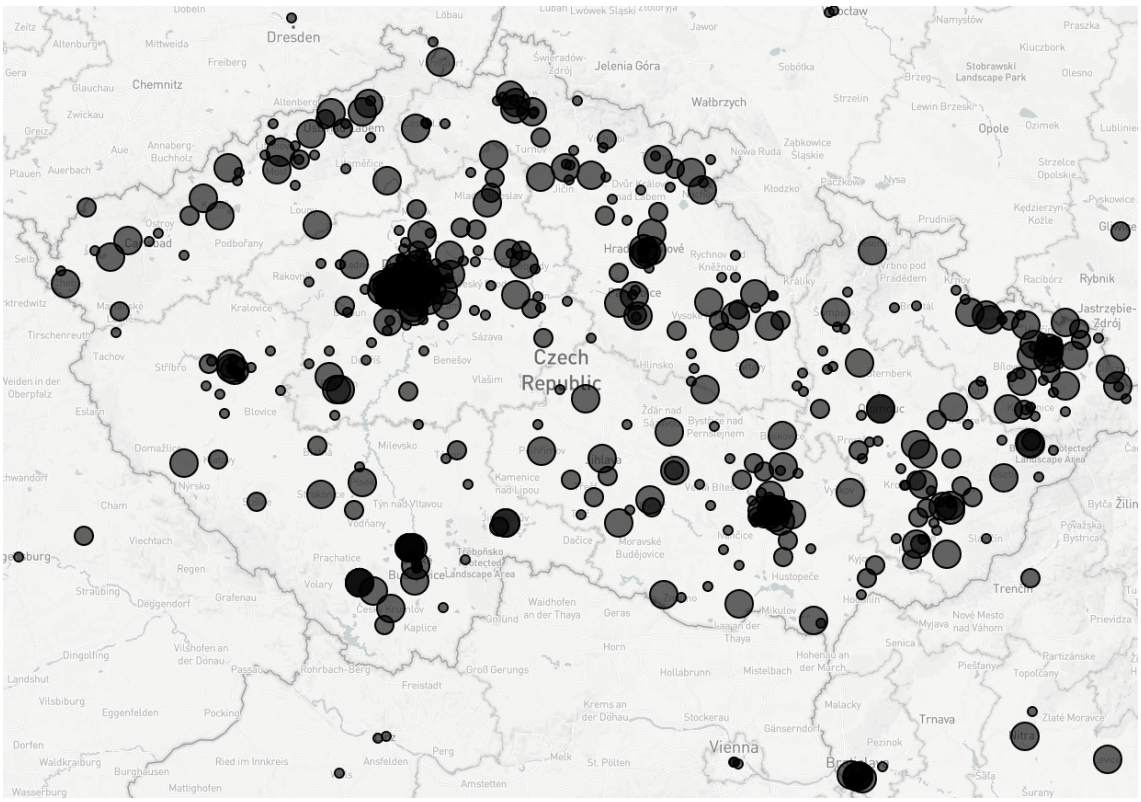
Autonomní systém	Počet hostovaných webů
AS15685 Casablanca INT	3322
AS43541 VSHosting s.r.o.	3168
AS24971 Master Internet s.r.o.	2860
AS197019 WEDOS Internet, a.s.	2123
AS24806 INTERNET CZ, a.s.	1134
AS39392 SuperNetwork s.r.o.	1013
AS25234 ACTIVE 24	945
AS5588 T-Mobile Czech Republic a.s.	590
AS34222 ZONER software, a.s	495
AS39790 Web4U s.r.o.	490
AS35236 2 connect a.s.	302
AS16276 OVH SAS	276
AS29134 Ignum s.r.o.	264
AS198171 HOSTING90 systems s.r.o.	237
AS35592 Coolhousing s.r.o.	228
AS8075 Microsoft Corporation	226
AS44984 Fortion Networks, s.r.o.	190
AS29208 Dial Telecom, a.s.	189
AS5610 O2 Czech Republic, a.s.	175
AS6830 Liberty Global Operations B.V.	130

4.5 Geografické umístění webových serverů

Geolokačním dotazem na umístění webových serverů obsluhujících webové stránky bylo zjišťováno umístění webových serverů. Tato data jsou vizualizována podle zeměpisných souřadnicích na mapě na obrázku 4.4. Tato mapa byla vygenerována skriptem `build_map.py` a je uložena v podobě interaktivní webové aplikace v souboru `vizualizace/coords.html`, který je součástí přílohy práce. Jednotlivé obrazce reprezentují dané zeměpisné místo, velikost těchto obrazců odpovídá počtu webových serverů v daném umístění.

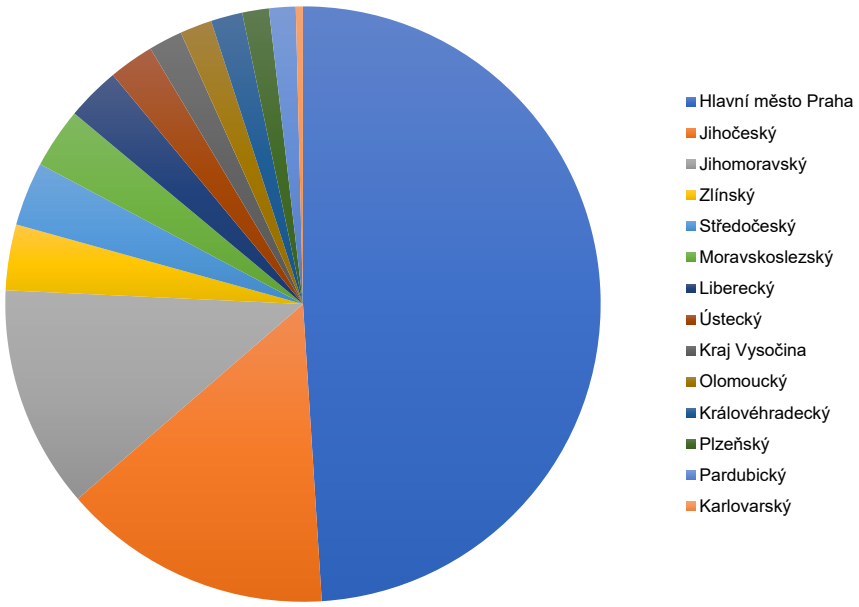


Obr. 4.3: Hostování webových stránek v autonomních systémech.

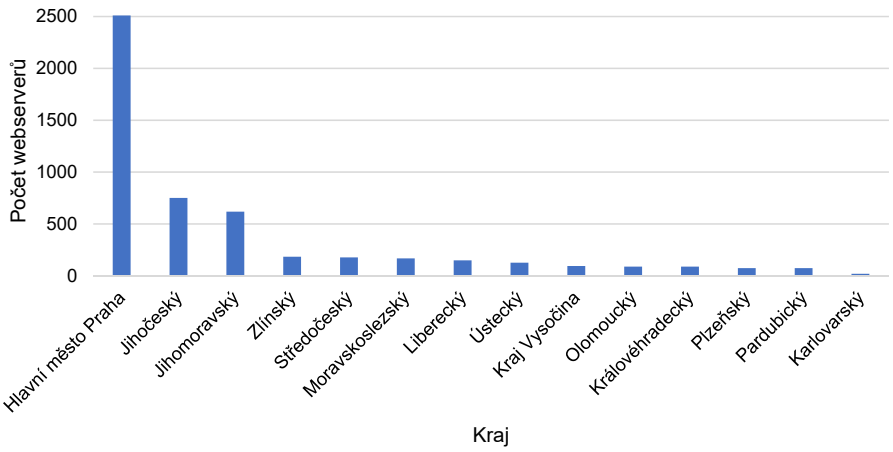


Obr. 4.4: Mapa geografického umístění detekovaných webových serverů. Autory mapových podkladů jsou přispěvatelé projektu OpenStreetMap, `http://www.openstreetmap.org`.

Tato data byla dále segmentována do grafu 4.6 zachycujícího počty detekovaných webových serverů v jednotlivých krajích České republiky, podíl zastoupení jednotlivých krajů zobrazuje graf 4.5.



Obr. 4.5: Podíl umístění webserverů v jednotlivých krajích ČR.



Obr. 4.6: Počet webserverů v jednotlivých krajích ČR.

Téměř polovina webů byla umístěna v Praze, velkou část tvoří také Jihomoravský kraj – tedy kraje, ve kterých jsou největší města republiky a také největší datacentra. Dále pak následují zastoupením v řádech jednotek procent další kraje. Výjimku tvoří Jihočeský kraj, v jehož městě Hluboká nad Vltavou sídlí datacentra jednoho z největších poskytovatelů webhostingových služeb v ČR, společnosti WEDOS. Tato

společnost se rozhodla pro vybudování svých datacenter na zeelené louce v menším městě a její vliv je velmi patrný i z těchto grafů.

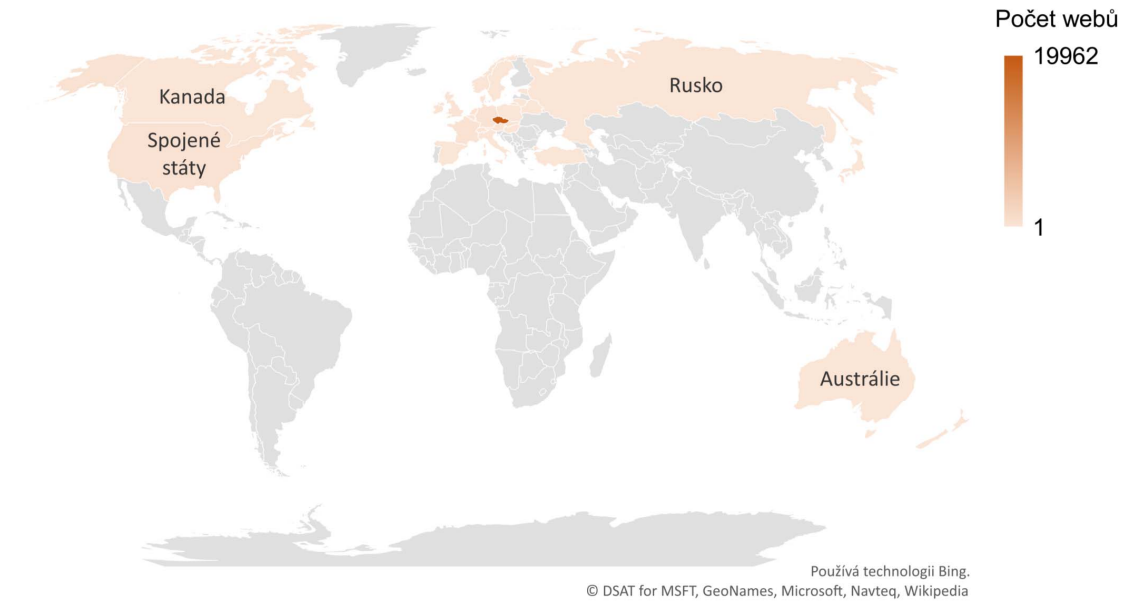
Údaje pochází z geolokačních databází a informace o poloze serverů jsou samozřejmě jen orientační, proto například na vizualizované mapě 4.4, v okolí zmíněné Hluboké nad Vltavou je více bodů, které všechny představují servery společnosti WEDOS. Tato organizace má přiděleno více sítí a geolokační databáze vrací pro každou síť rozdílnou polohu.

Některé webové stránky ze sledovaného vzorku dat nebyly hostovány na serverech v České republice. Nejvíce webových serverů mimo ČR bylo umístěno ve Spojených státech amerických. Z jisté části je to dáno využitím globálních cloudových služeb pro hostování webových aplikací, jako je například Amazon Web Services nebo Microsoft Azure.

Počet webů umístěných v jednotlivých zemích je uveden v tabulce 4.6. Země, ve kterých bylo detekováno hostování alespoň z jedné webové stránky jsou prezentovány na kartogramu 4.7.

Tab. 4.6: Umístění webových serverů podle zemí.

Země	**Počet**
Česká republika	19 962
USA	386
Francie	274
Německo	247
Slovensko	205
Velká Británie	95
Nizozemsko	91
Irsko	62
Rusko	37
Polsko	34
Itálie	31
Ostatní	70



Obr. 4.7: Mapa zemí hostujících sledované webové stránky.

4.6 Zastoupení webových serverů

Na základě informací z hlavičky HTTP odpovědi webových serverů bylo zjišťováno, jaký webový server obsluhuje konkrétní stránky. Toto zastoupení je uvedeno v tabulce 4.7. Celkové zastoupení webových serverů napříč všemi kategoriemi zobrazuje graf 4.8, zastoupení v rámci kategorií pak graf 4.9.

Tab. 4.7: Zastoupení webových serverů podle kategorií.

	Banky	Eshopy	Nemocnice	Pojišťovny	Reality	Řemesla	Státní organizace	Střední školy	Vysoké školy	Webhostingy
Apache	39%	50%	54%	54%	55%	53%	39%	62%	64%	62%
nginx	19%	31%	18%	8%	20%	19%	13%	12%	14%	31%
neuvedeno	31%	11%	16%	19%	12%	14%	22%	17%	17%	0%
Microsoft IIS	6%	6%	9%	12%	7%	11%	21%	7%	3%	5%
openresty	0%	1%	2%	0%	1%	1%	1%	1%	1%	3%
ostatní	6%	1%	0%	8%	2%	1%	5%	1%	2%	0%
lighttpd	0%	0%	1%	0%	3%	0%	0%	0%	0%	0%

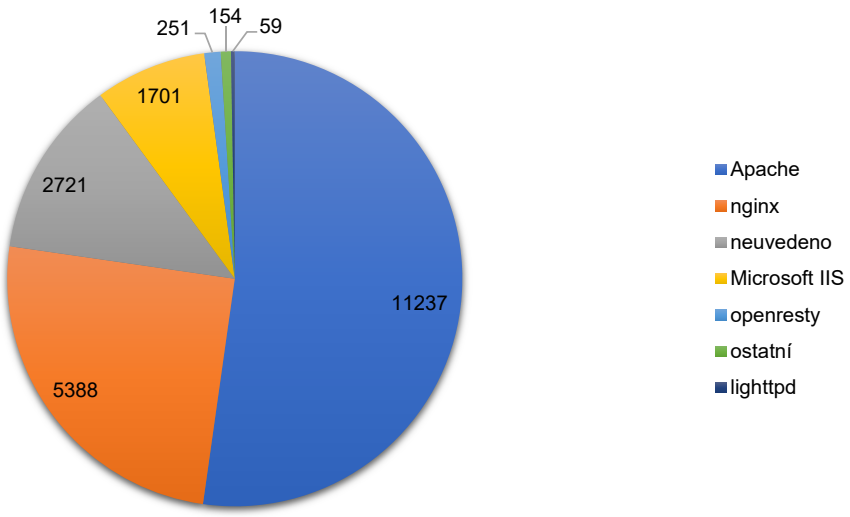
Výrazně nejpoužívanější webový server je webový server Apache. V provozu byla detekována celá řada jeho verzí. Nejstarší verze, která byla při prohledávání webů zaznamenána je 1.3.27, která byla vydána v říjnu 2002. Podle serveru CVE Details

[1]Zahrnuty pouze země s alespoň 30 detekovanými webservery.

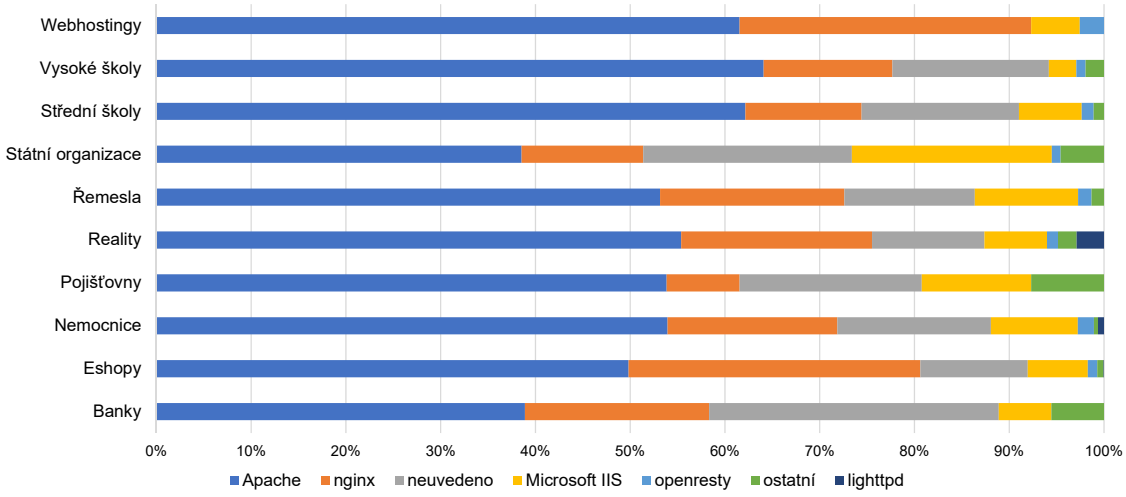
existuje pro tuto verzi celá řada velmi nebezpečných zranitelností [24] a z bezpečnostního hlediska rozhodně nelze použití takto zastaralé verze na produkčním serveru doporučit.

Nejvyšší zastoupení proprietárního serveru Microsoft Internet Information Services bylo zaznamenáno u webů státní správy. Naopak nejvíce proaktivní přístup otevřeným technologiím je patrný v resortu školství. V kategorii bankovnictví se v nejvyšší míře vyskytovaly různá řešení s vlastní signaturou, často založená na Java technologiích, které jsou ovšem díky nízkému výskytu zařazeny do kategorie ostatní. Zdaleka nejvyšší množství serverů neodeslalo hlavičku o použitém serveru právě v kategorii bank a státních organizací.

U kategorií s vysokým výskytem hostovaných stránek jsou díky použití shodných serverů výsledky téměř podobné a tato korelace svědčí platnost odhadů uvedených v podkapitole 4.2.



Obr. 4.8: Souhrnný podíl webových serverů.



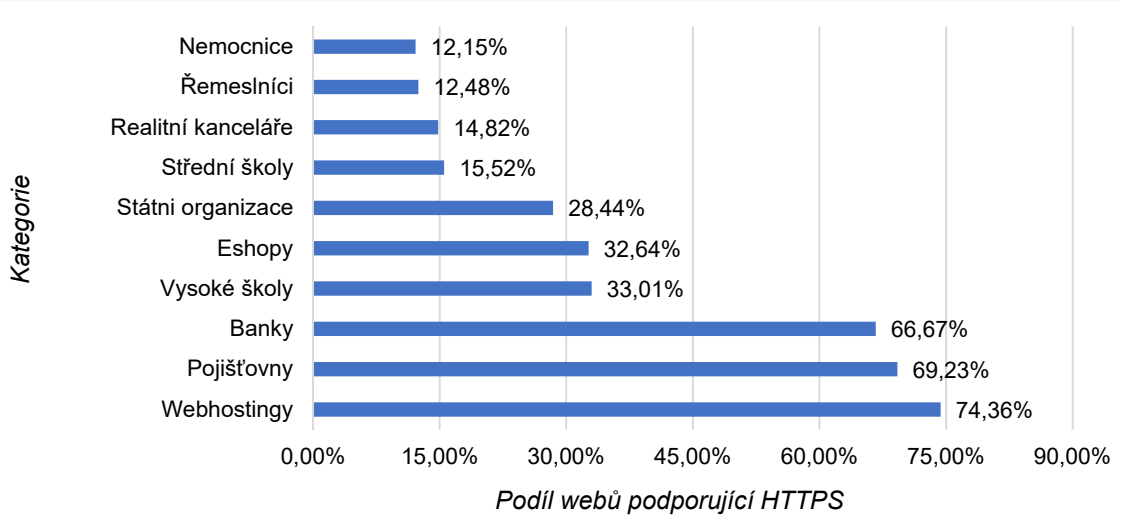
Obr. 4.9: Zastoupení webových serverů podle kategorií.

4.7 Podpora protokolu HTTPS webovými servery

Na základě chování webové stránky bylo zjišťováno, zda je korektně implementována podpora zabezpečeného protokolu HTTPS, včetně použití certifikátů, které testovací server považoval za důvěryhodné (byly součástí balíku `ca-certificates` linuxové distribuce Debian).

Některé weby fungují tak, že poskytují svůj obsah na obou protokolech, tedy jak HTTP, tak HTTPS současně. Jiné weby uživatele při příchodu na stránku pomocí protokolu HTTP přesměrují na tutéž stránku, avšak prostřednictvím protokolu HTTPS. Toto chování nebylo zohledňováno a vždy se detekční systém pokusil připojit jak prostřednictvím obou protokolů zvlášť.

Výsledky tohoto pozorování zobrazuje graf 4.10, protokol HTTPS podporovalo celkem 35,93 % sledovaných webů.



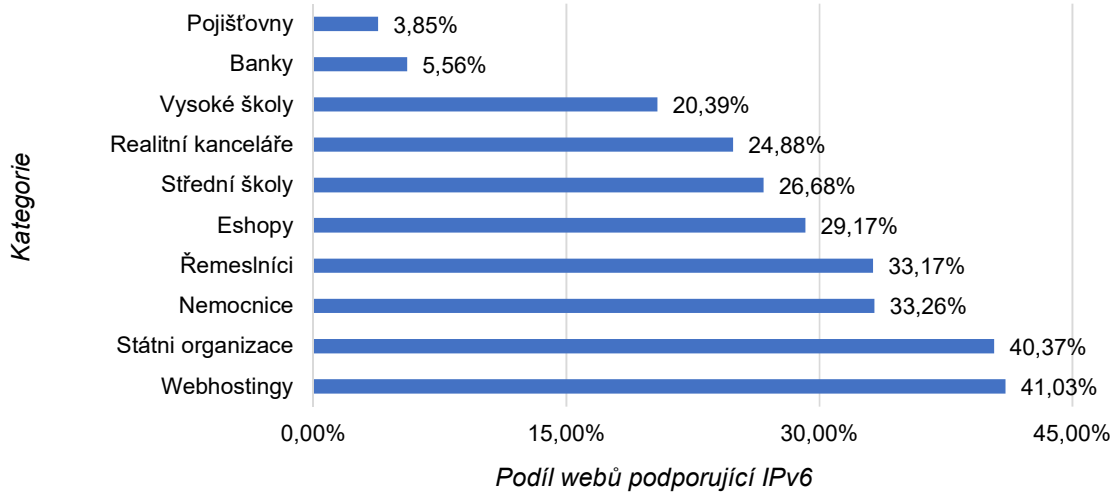
Obr. 4.10: Podíl webů přístupných protokolem HTTPS v jednotlivých kategoriích.

U této sledované vlastnosti je pozitivní, že zkoumané weby finančního sektoru tento protokol ve velké míře podporovaly, naopak překvapením byla poměrně nízká podpora ze strany webů vysokých škol.

4.8 Podpora protokolu IPv6 webovými servery

Během procházení webů jednotlivých kategorií subjektů byla testována podpora internetového protokolu IPv6. Všechny weby, které byly testovány poskytovaly svůj obsah také protokolem IPv4.

Podíl webů, jejichž webové servery jsou připojeny k internetu také prostřednictvím protokolu IPv6 a jejíž domény pomocí AAAA záznamů odkazují na příslušné IPv6 adresy je zobrazen v grafu 4.11. Jednalo se o celkem 25,83 % dotázaných webů.



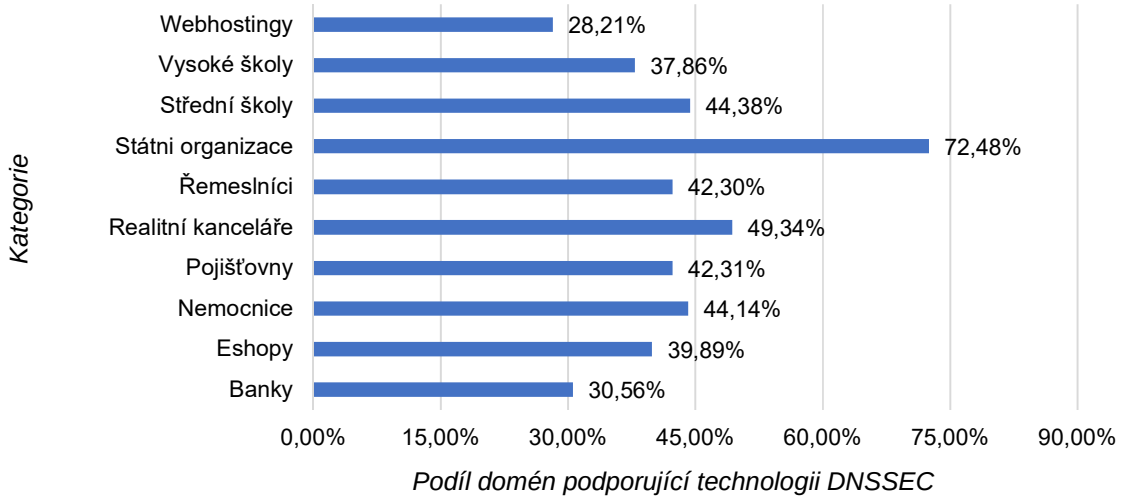
Obr. 4.11: Podíl webů přístupných protokolem IPv6 v jednotlivých kategoriích.

Značný podíl webů implementující tuto technologii byl mezi weby státních organizací. Jedná se o důsledek usnesení Vlády České Republiky [25], které stanoví za povinnost ústředním orgánům státní správy zajistit přístup k internetovým stránkám a poskytovaným veřejným službám státní služby současně protokoly IPv4 i IPv6 do konce roku 2010.

Nejméně webů, které nový internetový protokol podporují, je mezi komerčními bankami a pojišťovnami. Tento trend může být způsoben obecnou konzervativností těchto často nadnárodních korporací.

4.9 Podpora zabezpečení domény technologií DNSSEC

Během vyhodnocení navštívených stránek bylo testováno, zda daná stránka podporuje technologii DNSSEC, případně jakou sadu klíčů používá. Podíl webů, které implementují tuto technologii pro jednotlivé kategorie, zobrazuje graf 4.12.



Obr. 4.12: Podíl webů implementujících technologii DNSSEC podle kategorie.

Největší podíl webu implementujících tuto technologii byl zaznamenán v kategorii webů státní správy. Jedná se o důsledek usnesení Vlády České Republiky [26], které stanoví za povinnost ústředním orgánům státní správy zabezpečit všechny jimi držené domény technologií DNSSEC do června 2015.

U bankovních webů je možné naopak sledovat nejnižší zastoupení podílu podpory této technologie. Mnoho bankovních domů používá ale pro své systémy přímého bankovnictví zvláštní domény, které tuto technologii mohou mít implementovanou.

Nejčastější sady klíčů

V tabulce 4.8 je seznam nejčastěji se vykytovaných sad klíčů technologie DNSSEC na sledovaných doménách. Bylo ovšem zjištěno, že některé hostingy, z hlediska počtu výskytů, zejména FORPSI, využívá pro každý web zvláštní sadu klíčů, tedy proto nejsou tyto hostingy součástí tabulky.

Tab. 4.8: Nejčastější detekované sady klíčů DNSSEC.

Sada DNSSEC klíčů	**Počet výskytů**
A24-KEYSET	2036
WEDOS	1866
IGNUM	873
WEB4U	584
ONESOLUTION-KEYSET	373
KS:ZONER:1289219690	304
KEYSET	109
TELE3	93
ENDORA	67
BANAN.CZ-KEYSET	49
AEROHOSTING-CZ	46
REGISTRATOR	24
STABLECZ2	20
PK-CZ	18

5 ZÁVĚR

Tato diplomová práce se zabývala problematikou hostování webových stránek provozovaných na doménových jménech registrovaných v České Republice. Obsahem první části práce byl popis širší problematiky provozování webových stránek, byla popsána funkčnost systémů a principů důležitých pro zajištění přístupnosti webových stránek internetovým uživatelům. Zvláštní pozornost byla věnována specifikům této problematiky v České Republice.

Na základě těchto poznatků bylo navrženo několik přístupů k detekci hostování webových stránek na sdíleném serveru (tzv. webhosting). Tyto metody byly poté implementovány do aplikace zpracovávající zvolené webové stránky. Aplikaci tvoří kolekce modulů, které provádí sběr potřebných dat z webových stránek a příslušných registrů. Na základě těchto informací aplikace detekuje hostování dle popsaných metod. Realizovaná aplikace je modulární a obsahuje také moduly, které jsou použity pro automatizované generování vstupních souborů a provádění další analýzy nasbíraných dat. Základní funkčnost aplikace a jejich modulů je popsána a vyjádřena příslušnými diagramy.

Ve poslední části práce byla sestavena databáze více než 25 000 webů v deseti kategoriích. Jednalo se o kategorie internetových obchodů, komerčních bank, nemocnic, pojišťoven, provozovatelů webhostingu, realitních kanceláří, řemeslníků, státních organizací, středních škol a vysokých škol. Tyto weby byly použity jako vstup do realizované aplikace.

Získané výsledky jsou prezentovány pomocí komentovaných tabulek a grafů. Bylo zjištěno, že asi 97 % navštívených webů je umístěno na serverech provozovaných poskytovateli webhostingových služeb. V kategoriích, kde je menší zastoupení velkých organizacích je hostování webu na vlastním serveru prakticky ojedinělým jevem. Mimo to byly popsány další vlastnosti navštívených webových stránek, jako je podpora protokolů HTTPS, IPv6, technologie DNSSEC a také zastoupení webových serverů, které zpřístupňují tyto webové stránky. V analytické části práce jsou navíc prezentovány i další zjištěné skutečnosti o stavu kolem hostování webových stránek v České Republice.

Z těchto prezentovaných výsledků považuji za zajímavé srovnání podílů webů podporujících sledované technologie napříč sledovanými kategoriemi. Překvapila například obecně nízká podpora technologie podepisování DNS záznamů DNSSEC ze strany komerčních bank a pojišťoven. Zajímavé považuji také rozložení umístění webových serverů prezentované na obr. 4.7 a 4.4.

Součástí práce je příloha, která popisuje kroky nezbytné ke spuštění aplikace, popis jejího rozhraní a stručný návod k jejímu použití.

LITERATURA

[1] Brief History of the Internet. *Internet Society* [online]. [cit. 2016-11-15]. Dostupné z: `http://www.internetsociety.org/internet/what-internet/history-internet/brief-history-internet`

[2] RFC 1738: Uniform Resource Locators (URL) BERNERS-LEE, T. *The Internet Engineering Task Force* [online]. IETF, 1994[cit. 2016-11-1]. Dostupné z: `https://www.ietf.org/rfc/rfc1738.txt`

[3] RFC 1034: Domain Names - concepts and facilities. MOCKAPETRIS, P. *The Internet Engineering Task Force* [online]. IETF, 1987 [cit. 2016-11-1]. Dostupné z: `https://www.ietf.org/rfc/rfc1034.txt`

[4] RFC 791: Internet Protocol. POSTEL, Jon. *The Internet Engineering Task Force* [online]. IETF, 1981 [cit. 2016-11-15]. Dostupné z: `https://www.ietf.org/rfc/rfc0791.txt`

[5] RFC 2616: Hypertext Transfer Protocol – HTTP/1.1. FIELDING R. *The Internet Engineering Task Force* [online]. IETF, 1999 [cit. 2016-11-1]. Dostupné z: `https://tools.ietf.org/html/rfc2616`

[6] Hypertext Transfer Protocol – HTTP/1.1. *W3C* [online]. [cit. 2016-11-1]. Dostupné z: `https://www.w3.org/Protocols/rfc2616/rfc2616.html`

[7] PUŽMANOVÁ, Rita. *TCP/IP v kostce*. 2., upr. a rozš. vyd. České Budějovice: Kopp, 2009. 620 s. ISBN 978-80-7232-388-3.

[8] Number Resources. *IANA* [online]. [cit. 2016-12-2]. Dostupné z: `https://www.iana.org/numbers`

[9] AERTSEN, Maarten. *How to bring HTTPS to the masses?: Measuring issuance in the first year of Let's Encrypt* [online]. 2016, , strany 21-27 [cit. 2017-05-07]. Dostupné z: `https://www.sidnlabs.nl/downloads/theses/How-to-bring-HTTPS-to-the-masses_measuring-1y-of-LE.pdf`

[10] KABELOVA, Alena. *DNS in Action: A Detailed and Practical Guide to DNS Implementation, Configuration, and Administration*. Olton: Packt Publishing, 2006. ISBN 9781904811787.

[11] HÁLA, Tomáš. DNSSEC - aktuální stav a zkušenosti z praxe. *CZ.NIC, z. s. p. o.* [online]. 2010 [cit. 2017-03-18]. Dostupné z: `https://www.nic.cz/public_media/IT10/prezentace/den_1_8_Hala.pdf`

[12] Apache Virtual Host documentation. *The Apache Software Foundation* [online]. [cit. 2016-12-2]. Dostupné z: `https://httpd.apache.org/docs/2.4/vhosts/`

[13] February 2016 Web Server Survey. *NETCRAFT* [online]. 2016 [cit. 2016-11-24]. Dostupné z: `https://news.netcraft.com/archives/2016/02/22/february-2016-web-server-survey.html`

[14] Pravidla registrace doménových jmen v ccTLD .cz. *CZ.NIC* [online]. CZ.NIC, 2007 [cit. 2016-11-3]. Dostupné z: `https://www.nic.cz/files/nic/doc/pravidla_registrace_CZ.pdf`

[15] CHAPPELL, Laura a TITTEL, Ed. *Guide to TCP/IP*. 3rd ed. Boston, Mass: Thompson/Course Technology, 2007, 742 s. ISBN 1418837555.

[16] MILLER, Philip M. *TCP/IP: The ultimate protocol guide: Volume 2 - Applications, Access and Data Security*. Boca Raton: BrownWalker Press, 2009. ISBN 978-1-59942-493-4.

[17] Hosting. *CZ.NIC* [online]. CZ.NIC, 2016 [cit. 2016-11-25]. Dostupné z: `https://stats.nic.cz/stats/hosting/`

[18] PILGRIM, M. *Ponořme se do Python(u) 3*. CZ.NIC, 2010. 435 s. ISBN: 978-80-904248-2-1.

[19] HOMAN, Jacqueline. *Relational vs. non-relational databases: Which one is right for you?* [online]. Pluralsight, 2014 [cit. 2017-04-28]. Dostupné z: `https://www.pluralsight.com/blog/software-development/relational-non-relational-databases`

[20] Features Of SQLite. *SQLite* [online]. SQLite Team, 2002 [cit. 2017-03-27]. Dostupné z: `https://www.sqlite.org/features.html`

[21] Sorting and Searching. SKIENA, Steven. *The algorithm design manual*. 2nd ed. London: Springer, 2008, s. 103-144. ISBN 9781848000704. Dostupné také z: `http://link.springer.com/chapter/10.1007%2F978-1-84800-070-4_4`

[22] *Adresář VŠ, přímo řízených organizací MŠMT a orgánů státní správy a samosprávy* [online]. Praha: Ministerstvo školství, mládeže a tělovýchovy, 2017 [cit. 2017-04-03]. Dostupné z: `http://stistko.uiv.cz/proavs/provsass.asp`

[23] BULÍN, Martin. *Analýza realitního trhu pomocí informací na Internetu*. Brno, 2017. 75 s. Diplomová práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací.
Vedoucí práce: Doc. Ing. Dan Komosný, Ph.D.

[24] Apache 1.3.27 *Security Vulnerabilities* [online]. CVE Details, 2017 [cit. 2017-05-16] Dostupné z: https://www.cvedetails.com/vulnerability-list/vendor_id-45/product_id-66/version_id-8205/Apache-Http-Server-1.3.27.html

[25] *USNESENÍ VLÁDY ČESKÉ REPUBLIKY ze dne 8. června 2009 ke Zprávě o přechodu na internetový protokol verze 6 (IPv6)*. Praha: Vláda České Republiky, 2009, číslo 727.

[26] *USNESENÍ VLÁDY ČESKÉ REPUBLIKY ze dne 18. prosince 2013 ke Zprávě o zavádění technologie DNSSEC a o plnění usnesení vlády ze dne 8. června 2009 č. 727, ke Zprávě o přechodu na internetový protokol verze 6 (IPv6)*. Praha: Vláda České Republiky, 2013, číslo 982.

[27] Python 3.6.1 Documentation: The Python Standard Library. *Python 3.6.1 Documentation* [online]. Python Software Foundation, 2017 [cit. 2017-05-02]. Dostupné z: https://docs.python.org/3/library/

[28] Python 3.6.1 Documentation: Creation of virtual environments. *Python 3.6.1 Documentation* [online]. Python Software Foundation, 2017 [cit. 2017-05-02]. Dostupné z: https://docs.python.org/3/library/venv.html

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

API Application Programming Interface

CSV Comma-separed Values

DNS Domain Name System

FIFO First In, first Out

FQDN Fully Qualified Domain Name

FTP File Transfer Protocol

FTPS File Transfer Protocol Secure

FUP Fair Usage Policy

HTML Hypertext Markup Language

HTTP Hypertext Transfer Protocol

HTTPS Hypertext Transfer Protocol Secure

IANA Internet Assigned Numbers Authority

IIS Internet Information Server

IPv4 Internet Protocol version 4

IPv6 Internet Protocol version 6

ISP Internet Service Provider

PKI Public Key Infrastructure

RIR Resource Internet registry

RIPE NNC Réseaux IP Européens Network Coordination Centre

SCP Secure Copy

SQL Structured Query Language

SŘBD Systém řízení báze dat

SSH Secure Shell

SSL Secure Sockets Layer

WYSIWYG What you see is what you get

TCP Transmission Control Protocol

TLD Top Level Domain

TLS Transport Layer Security

TTL Time To Live

URI Uniform Resource Identifier

URL Uniform Resource Locator

VPS Virtuální privátní server

WWW World Wide Web

SEZNAM PŘÍLOH

A OBSAH PŘILOŽENÉHO CD

Obsahem přiloženého CD jsou zdrojové realizované soubory aplikace, elektronická verze textu práce a souhrnné výsledky realizované analýzy. Soubory na přiloženém CD jsou organizovány do následující adresářové struktury.

- `/` ..... Kořenový adresář přiloženého CD
  - `evaluation` ..... Zdrojové soubory realizované aplikace
  - `examples` ..... Příklady použití aplikace
    - `source_files` ..... Adresář pro umístění vstupních souborů
    - `visualisation` ..... Adresář pro umístění generovaných map
  - `sql_queries` ..... SQL dotazy užité pro zpracování analýzy
  - `thesis` ..... Elektronická podoba této práce

B APLIKACE PRO DETEKCI HOSTOVÁNÍ WEBOVÝCH STRÁNEK

Tato příloha popisuje potřebné kroky nutné ke přípravě prostředí, které umožňuje spuštění realizované aplikace pro získávání a analýzu dat ze vstupního souboru. Zdrojové soubory aplikace jsou k dispozici na přiloženém CD nosiči. Vzhledem k tomu, že Python je interpretovaný jazyk, není nutné před spuštěním aplikace zdrojové kódy kompilovat [18].

Upozornění. Aplikace se během své činnosti připojuje k dalším webovým službám a registrům. Uživatel, který spouští tuto aplikaci musí být srozuměn s podmínkami použití těchto služeb a aplikaci používat v souladu s těmito podmínkami. Aplikace si pro účel vyhodnocení uchovává v rámci vytvořené SQLite databáze informace, které pochází z těchto zdrojů.

Jmenovitě se jedná o tyto služby:

- WHOIS registr sdružení CZ.NIC (`https://www.nic.cz/whois/`),
- WHOIS registry regionálních registrátorů podřízených organizaci IANA (`http://www.iana.org/whois`),
- službu IP-API (`http://ip-api.com/`),
- a službu MapBox.org (`https://www.mapbox.com/`), která se využívá pouze při generování interaktivních map.

B.1 Požadavky na hostitelský systém

Protože je aplikace realizována v jazyce Python, který je multiplatformní, může být provozována na různých operačních systémech. Aplikace byla autorem vyvíjena a testována na distribuci Fedora 25 operačního systému Linux s použitým jádrem verze 4.10 a instalovaným interpretem jazyka Python 3.5.2.

Obecně jsou však pro korektní běh aplikace a možnosti spuštění všech implementovaných detekčních funkcionalit vyžadovány následující požadavky na hostitelský systém:

- Vzhledem k některým použitým konstrukcím jazyka Python je minimální požadovaná verze interpreta Python 3.4 [18],
- hostitel musí mít funkční konektivitu IPv4 i IPv6,
- hostitel a používaný DNS resolver musí podporovat překlad A záznamů (na IPv4 adresy), tak i AAAA záznamů (na IPv6 adresy) pro dotazované domény,
- v systému musí být aktuální certifikáty kořenových důvěryhodných autorit (CA),
- součástí Python distribuce musí být požadované balíčky (viz dále).

Z těchto důvodů není součástí souborů na přiloženém CD nosiči obsah této databáze, vstupních a výstupních souborů použitých pro vypracování analýzy. Součástí přiložené aplikace jsou ovšem veškeré nástroje, kterými je možné všechny tyto údaje sestavit.

Požadované závislosti

Aplikace pro svoji činnost využívá několika specializovaných knihoven pro práci se vzdálenými zdroji. Požadované používané balíčky, které nejsou součástí standardní knihovny Pythonu, jsou uvedeny v tabulce B.1. Kromě těchto balíčků jsou použity také některé, které tvoří součást standardní knihovny Pythonu [27], například balíček `sqlite3` pro zajištění práce s vestavěnou relační databází SQLite.

Tab. B.1: Python balíčky použité v aplikaci.

Balíček	Testovaná verze	Domovská stránka projektu
dnspython	1.15.0	http://www.dnspython.org/
folium	0.3.0	https://github.com/python-visualization/folium
ipwhois	0.15.1	https://github.com/secynic/ipwhois
pyquery	1.2.17	https://github.com/gawel/pyquery
requests	2.14.2	http://www.python-requests.org
tldextract	2.0.2	https://github.com/john-kurkowski/tldextract

B.2 Příprava prostředí pro běh aplikace

V následující části je popsána příprava prostředí pro spuštění aplikace v čistě nainstalovaném systému Ubuntu 16.04 (pro účely popisu bez grafického prostředí, předvedeno v rámci Windows Subsystem for Unix).

1. Ve výchozím stavu je již interpret Python 3.5 přítomný, to je možné ověřit příkazem:

   ```
   $ python3 --version
   ```

   V případě, že by příkaz nebyl vykonán a interpret není nainstalován, je možné nainstalovat Python 3.5 příkazem `sudo apt install python3`.
2. Pro instalaci balíčků je možné doinstalovat nástroj `pip` pro snadnou instalaci Python balíčků z repositáře.

   ```
   $ sudo apt install python3-pip
   ```

Spolu s tímto balíkem budou nainstalovány některé další vývojové knihovny a kompilátory, pokud již v systému nejsou přítomny.

3. Pomocí nástroje `pip` je možné nainstalovat požadované další chybějící balíky:

   ```
   $ sudo pip3 install pyquery tldextract folium
     dnspython requests ipwhois
   ```

4. Pomocí nástroje `pip` je vhodné zaktualizovat balíček `requests`, který je součástí distribuce Pythonu 3 v distribuci Ubuntu na nejnovější verzi:

   ```
   $ sudo pip install --upgrade requests
   ```

5. Tímto by měly být nainstalovány všechny potřebné závislosti, to je možné ověřit například zavoláním skriptu `process_csv.py` s parametrem pro zobrazení nápovědy:

   ```
   $ ./process_csv.py --help
   ```

Poznámka. Z důvodů načítání modulů je vyžadováno, aby všechny skripty byly spuštěny v pracovním adresáři, ve kterém se nachází složka `modules`.

Využití virtuálního prostředí virtualenv

Pokud nechcete instalovat dodatečné balíčky do systémového umístění, je možné volitelně využít virtuální prostředí příkazem `venv`, který zajistí vytvoření takového prostředí přímo v pracovním adresáři izolovaným od systémových složek. Každé takové prostředí má vlastní binární komponenty interpretu Python [28].

Ukázka vytvoření virtuálního prostředí pro aplikaci při použití shellu `bash`, jeho aktivace a následné deaktivace:

```
$ python3 -m venv /home/demo/eval/environment
$ source /home/demo/eval/environment/bin/activate
(env) $ == práce v~prostředí ==
(env) $ deactivate
$
```

Vytvoření databázového schématu aplikace

Pro vytvoření požadovaného schématu databáze je nutné v databázi spustit následující příkazy:

```
CREATE TABLE "results" (
"domain"  VARCHAR(200),
"timestamp"  INTEGER DEFAULT CURRENT_TIMESTAMP,
"hosted_rev"  NUMERIC,
"hosted_known"  NUMERIC,
"hosted_whois"  NUMERIC,
"hosted_email"  NUMERIC,
"server_httpd"  VARCHAR(200),
"server_xpowerer"  VARCHAR(50),
"server_type"  VARCHAR(50),
"server_ip"  VARCHAR(16),
"server_ip_name"  TEXT(255),
"server_ip_domain"  TEXT(255),
"dnssec_keyset"  TEXT(100),
"domain_holdername"  TEXT(150),
"ip_holdername"  TEXT(150),
"category"  VARCHAR(25),
"original_url"  TEXT,
"note"  TEXT,
"https_support"  NUMERIC,
"ipv6_addr"  TEXT,
PRIMARY KEY ("domain" ASC),
CONSTRAINT "FK_results_wevserver" FOREIGN KEY ("server_ip")
REFERENCES "webserver" ("ip") ON DELETE RESTRICT
ON UPDATE RESTRICT
);

CREATE TABLE "webhoster" (
`hoster_domain`        VARCHAR(255),
`company`        VARCHAR(100),
PRIMARY KEY(hoster_domain)
);

CREATE TABLE "webserver" (
`ip`        VARCHAR(30),
```

```
`timestamp`        TIMESTAMP NOT NULL DEFAULT CURRENT_TIMESTAMP,
`as`         VARCHAR(140),
`organization`        VARCHAR(100),
`isp`         VARCHAR(100),
`country`         VARCHAR(10),
`lat`         NUMERIC,
`lon`         NUMERIC,
`city`         INTEGER,
`region`         VARCHAR(40),
`postal_code`         VARCHAR(40),
`contact_email`         VARCHAR(255),
`contact_address`         TEXT,
PRIMARY KEY(`ip`)
);
```

Tato SQL věta je přiložena v adresářové struktuře přiloženého CD v souboru `eval_schema.sql` a pro její spuštění může být využit skript `database.py`, s následujícími parametry:

```
$ ./database.py init
```

C PARAMETRY ROZHRANÍ APLIKACE

Vzhledem k modulárnímu způsobu návrhu aplikace, který umožnil rozdělení jednotlivých funkčních celků aplikace a pro uživatele představuje přínos v možnosti využití jen konkrétní požadované části funkcionality aplikace, disponuje aplikace několika spustitelnými skripty v interpretovaném jazyce Python 3. Veškeré tyto skripty implementují argument `--help`, zkráceně `-h`, který vypíše nápovědu se seznamem přijímaných parametrů. Jednotlivé skripty mohou být volány příkazem např. `./<název skriptu>.py` z pracovního adresáře, ve kterém je přítomný adresář s pomocnými moduly aplikace `modules`. Pro použití tohoto způsobu volání je nutné v unixových operačních systémech nastavit těmto skriptům oprávnění pro spouštění (například příkazem `chmod +x <název skriptu>.py`).

Následuje výpis přijímaných argumentů jednotlivých spustitelných skriptů. Argument pro výpis nápovědy již není dále uváděn. Vstupně/výstupní rozhraní včetně vestavěných nápověd komunikuje s uživatelem v anglickém jazyce. Pro účely tohoto popisu jsou níže tyto informace přeloženy do češtiny.

Generování seznamu webů z katalogu Firmy.cz

použití: `build_db_firmy.py [-h] [--starting NUM] category outputfile`
povinné argumenty:

```
category          Vybraná kategorie katalogu Firmy.cz
outputfile        Název výstupního souboru
```

volitelné argumenty:

```
--starting NUM    Specifikace první stránky
```

Generování seznamu webů z katalogu Toplist.cz

použití: `build_db_toplist.py [-h] [--starting NUM] category outputfile`
povinné argumenty:

```
category          Vybraná kategorie katalogu Toplist
outputfile        Název výstupního souboru
```

volitelné argumenty:

```
--starting NUM    Specifikace prvního záznamu katalogu
--delimiter DELIM  Definovat oddělovač CSV souboru
```

Vykreslení mapy serverů

Modul vykreslí interaktivní mapu všech detekovaných webserverů do složky `visualisation`.
použití: `build_map.py [-h] {coords}`

Export výsledků do souboru

Skriptem je možné vyexportovat výsledky vybrané kategorie do CSV souboru, název výstupního souboru odpovídá tvaru `<kategorie>_export.csv`.
použití: `export_csv.py [-h] category`
povinné argumenty:

```
category          Vybraná kategorie pro export
```

Výpis výsledků kategorie

Skriptem je možné vypsat výsledky vybrané kategorie na systémový standardní výstup.
použití: `print_category.py [-h] category`
povinné argumenty:

```
category          Vybraná kategorie pro výpis
```

Výpis výsledků kategorií

Skriptem je možné vypsat souhrnný přehled výsledků metod a počtů zpracovaných stránek podle jednotlivých kategorií.
použití: `print_stats.py [-h]`

Výpis webů hostovaných na webovém serveru nebo webhostingu.

Na základě vyhodnocených webů v databázi vypíše nalezené domény, které byly umístěny na konkrétním webovém serveru. Webový server musí být specifikován pomocí IPv4 adresy konkrétního serveru případně specifikovaný pomocí adresy sítě ve formátu `X.X.X.X/YY`, kde `YY` určuje dekadicky počet bitů, které tvoří adresu sítě. Podporovány jsou pouze hodnoty 24, 16 a 8, tedy síťové masky `255.255.255.0`, `255.255.0.0` a `255.0.0.0`.
použití:

```
print_pages.py [-h] [--summary] [--webhoster] QUERY
```

povinné argumenty:

```
QUERY          Adresa serveru nebo sítě webhostera,
               případně doménové jméno webhostera
```

volitelné argumenty:

```
--summary      Vytiskne statistiku hostování webů
               zvoleného webhostera
--webhoster    Režim vyhledávání podle doménového jména
               webhostera
```

Hromadné vyhodnocení vstupního CSV souboru

Zpracuje zadaný CSV soubor a výsledky uloží do databáze podle zadané kategorie.
použití:

```
process_csv.py [-h] [--notecolumn N] [--noterow]
[--delimiter X] [--quotechar Q] [--skip S]
[--restoreconn] [--reevaluate] [--noczommit]
[--nosubdomommit] CAT filename N
```

povinné argumenty:

```
CAT                Volba kategorie vstupu
filename           Umístění vstup. souboru
N                  Pořadí sloupce s  URL
```

volitelné argumenty:

```
--notecolumn N     Specifikace sloupce s poznámkou
--noterow          Celý vstupní řádek se uloží do pozn.
--delimiter X      Oddělovač sloupce v CSV souboru
--quotechar Q      Spojovník dat v CSV souboru
--skip S           Přeskočit prvních S řádků
--restoreconn      Obnovovat připojení k WHOIS
--reevaluate       Přepsat výsledky v DB
--noczommit        Nepřeskakovat weby mimo .cz
--nosubdomommit    Nepřeskakovat subdomény
```

Hromadné vyhodnocení vstupního CSV souboru

Zpracuje pouze jednu zadanou doménu a uloží ji do databáze podle zadané kategorie.
použití:

```
process_page.py [-h] [--restoreconn] [--reevaluate]
[--noczommit] CAT DOMAIN
```

povinné argumenty:

```
CAT                 Volba kategorie vstupu
DOMAIN              Doména ke zpracování
```

volitelné argumenty:

```
--restoreconn       Obnovovat připojení k WHOIS
--reevaluate        Přepsat výsledky v DB
--noczommit         Nepřeskakovat weby mimo .cz
```

Pomocné moduly

Údržba databáze

Umožňuje provádět údržbu databáze. Implementuje následující operace:

- `addhoster` - přidá webhostera do databáze známých hostingů,
- `clean` - vyčistí zvolenou databázovou tabulku,
- `init` - vytvoří databázový soubor a jeho schema,
- `runscript` - spustí zadaný SQL dotaz.

použití:

```
modules/database.py [-h]
{addhoster,clean,init,runscript} SUB
```

povinné argumenty:

```
{addhoster,clean,init,runscript}
                        Požadovaná operace
SUB                     Argument požadavku
```

Zobrazení informací o serveru

Zobrazí zjištěných informací o softwarovém vybavení webového serveru obsluhujícím danou doménu.
použití:

```
modules/webserver_stats.py [-h] [-s] URL
```

povinné argumenty:

```
URL   URL požadavku k zaslání
```

volitelné argumenty:

```
-s    Vynutí použití protokolu HTTPS
```

Zobrazení výsledků WHOIS požadavku

Zobrazí zjištěných informací z WHOIS o požadovaném objektu. Zajišťuje zjištění následujících typů požadavků:

- `contact` - informace o kontaktu v .cz zóně reprezentované identifikátorem
- `domain` - informace o doméně v .cz zóně reprezentované identifikátorem
- `ip` - informace o IP adrese

použití:

```
modules/whois_parser.py [-h] [--restoreconn] TYPE SUBJECT
```

povinné argumenty:

```
TYPE      Typ objektu
SUBJECT   Identifikátor požadovaného subjektu
```

volitelné argumenty:

```
-restoreconn    Spustí uživatelský skript pro obnovu
                připojení při nedostupnosti registru
```

Příklad použití

Pro získání seznamu cestovních kancelářích z katalogu Firmy.cz, uložení tohoto seznamu do připraveného adresáře `source_files`, jeho vyhodnocení (s výchozími parametry) a vypsání výsledků vyhodnocení této kategorie je možné použít následující příkazy. Více příkladů je uvedeno na přiloženém CD v adresáři `examples`.

```
$ ./build_db_firmy.py Cestovni-sluzby/Cestovni-kancelare-a
-agentury source_files/cestovky.txt
$ ./process_csv.py cestovky source_files/cestovky.txt 1
$ ./print_category.py cestovky
```